UNIVERSIDADE DE LISBOA FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Aplicação de um Sistema Multiclassificador baseado na Divergência de Kullback-Leibler para Produção Automática de Mapas de Uso/Ocupação do Solo com base em Imagens de Satélite

Joel Dinis Baptista Ferreira da Silva

Orientadores:
Doutor Hugo Miguel Saiote Carrão
Doutora Ana Navarro

MESTRADO EM ENGENHARIA GEOGRÁFICA

2010

Índice

Resumo

O presente trabalho foi desenvolvido no contexto do projecto *DesertWatch-Extention*, que visa o desenvolvimento de uma aplicação informática que permita a derivação de indicadores de desertificação a partir de imagens de satélite. Um dos parâmetros de entrada para a determinação desses indicadores é um mapa de uso/ocupação de solo da área em estudo. Estes tipos de mapas têm servido de ponto de partida para o estudo do fenómeno da desertificação desde da década de 90, altura em que o uso de imagens de satélite para o estudo e gestão da superfície terreste ganhou o seu maior ímpeto.

O objectivo deste trabalho é, então, definir um algoritmo de classificação de imagens de satélite que seja simples, em termos de implementação e de aplicação pelos futuros utilizadores, mas também que possibilite a elaboração de mapas de uso/ocupação de solo com qualidade temática de, pelo menos, 80%. Nestas condições, foram seleccionados cinco classificadores já amplamente estudados e aplicados em problemas de classificação: o classificador discriminante linear, o classificador discriminante quadrático, o classificador de Parzen, o classificador k nearest neighbor e as árvores de classificação. Para uma comparação objectiva destes algoritmos, foram definidos quatro factores que ponderaram a selecção do classificador, são eles: as exigências computacionais, a parametrização, o volume de treino, a robustez ao ruído e a qualidade temática do mapa produzido. Os resultados mostram diferenças compatíveis com o que é afirmado na literatura. Em particular, os resultados mostram que, em termos de qualidade temática, os mapas resultantes com cada um dos algoritmos são equivalentes. Contudo, alguns erros de classificação persistem, como é o caso da confusão entre as classes de Agricultura de Sequeiro e Pastagens. Assim, procurou-se melhorar os resultados obtidos melhorando um dos algoritmos de classificação. O algoritmo seleccionado para esta finalidade foi o classificador linear discriminante.

Para refinar este algoritmo, adoptou-se um sistema de classificador composto, ou multiclassificador, do tipo cooperativo. E de modo a aumentar a separabilidade entre classes, foi introduzida no algoritmo a medida de entropia relativa, conhecida também por divergência de Kullback-Leibler, com a finalidade de identificar as classes mais prováveis de confusão com um dada classe inicial, reduzido a lista de classes possíveis a serem atribuídas a um determinado *pixel* Finalmente, o algoritmo determina uma transformação de Fisher, baseada nas classes previamente identificadas, que irá projectar o espaço de classificação inicial para outro espaço mas máxima separabilidade entre as classes, sendo a classificação final realizada nesse espaço transformado. Os resultados mostram um ganho significativo em exactidão temática global com apenas um pequeno ganho em tempo de classificação.

Palavras-Chave: Classificação automática, imagens de satélite, mapas de uso/ocupação de solo, divergência de Kullback-Leibler.

Abstract

This work was developed under the project DesertWatch-Extention, which aims to develop a computer application that allows the derivation of indicators of desertification from satellite images. One of input parameters for the determination of these indicators is a map of land use / land cover of the study area. These types of maps have served as a starting point for the study of desertification since the 90s, when the use of satellite imagery for the study and management of terrestrial surface has gained its greatest impetus.

The aim of this work is then set a classification algorithm of satellite imagery that is simple in terms of implementation and application by future users, but also allowing for mapping land use / land cover with a thematic quality of at least 80%. Accordingly, we selected five classifiers already widely studied and applied in classification problems: the linear discriminant classifier, a quadratic discriminant classifier, the Parzen classifier, the classifier k nearest neighbor and classification trees. For an objective comparison of these algorithms, we defined four factors that considered the selection of the classifier, they are: the computational requirements, the parameterization, the volume of training, the robustness to noise data and quality of the thematic map produced. The results show differences consistent with what is stated in the literature. In particular, the results show that, in terms of thematic quality, the maps derived with each of the algorithms are equivalent. However, some misclassification persist, as is the case of confusion between the classes of Rainfed Agriculture and Pasture. Thus, we tried to improve the results improving the classification algorithms. The algorithm chosen, for this proporse, was the linear discriminant classifier.

To refine this algorithm, it was adopted a cooperative multiclassifier strategy. And in order to increase the separability between classes, was introduced in the algorithm the measure of relative entropy, also known

as Kullback-Leibler divergence, with the aim of identifying the classes most likely to be confused with a given initial class, reducing, in this manner, the list of possible classes to be assigned to a given pixel Finally, the algorithm determines a transformation of Fisher, based on classes previously identified, which will project the initial classification space to another space but with maximum separability between classes, being the final classification made in this transformed space. The results show a significant gain in overall thematic accuracy with only a small gain in classification time.

Keywords: Automatic classification, satellite imagery, land use / land cover maps, Kullback-Leibler divergence.

Agradecimentos

Gostaria de agradecer à equipa do Grupo de Detecção Remota do Instituto Geográfico Português pelo apoio durante a realização deste trabalho com o qual aprendi e ainda aprendo bastante. Gostaria também de agradecer aos meus orientadores pela sua disponibilidade e interesse no trabalho que foi sendo realizado. Em particular, gostaria de agradecer ao Dr. Hugo Carrão que, de certa forma, me serviu de modelo na forma de abordar as questões encontradas durante a execução deste trabalho. Por fim, gostaria de agradecer à minha mãe e irmã, que desde sempre me apoiaram em qualquer situação. A todos o meu sincero obrigado.

Índice de Figuras

Índice de Tabelas

Índice de Quadros

Lista de Acrónimos

CST - *Commitee on Science and Technology*

DUE - *Data User Element*

DWE - DesertWatch Extention

DWO - DesertWatch Original

ESA - European Spacial Agency

FDA - *Fisher Discriminant Analysis*

IGP - Instituto Geográfico Português

IST CERENA - Instituto Superior Técnico, Centro de Recursos Naturais e Ambiente

k-NN - *k Nearest Neighbor*

LCCS - *Land Cover Classification System*

LDC - *Linear Discriminant Classifier*

LULC - *Land Use / Land Cover*

NDVI - N*ormalized Difference Vegetation Index*

QDC - *Quadratic Discriminant Classifier*

UNCCD - *United Nations Convention to Combat Desertification*

1 Introdução

Nesta secção é a apresentado o contexto em que decorreu o presente estudo (secção 1.1. Enquadramento). Procura-se mostrar as principais motivações que definiram as questões que o presente estudo procura responder. Na secção 1.2. procurou-se resumir os conceitos mais genéricos sobre o processo de classificação de imagens de satélite, assim como as abordagens já exploradas nesta área.

1.1 Enquadramento

A detecção remota tem sido utilizada para monitorizar os efeitos da desertificação e como informação base para a estimação de indicadores, por meio da classificação de imagens de satélite (Lantieri, 2003). A aplicação da detecção remota na monitorização do fenómeno da desertificação tem o seu início na década de 90, mostrando a sua utilidade na construção de séries temporais de indicadores desse fenómeno (Panigada et al., 2009). Os indicadores de desertificação são derivados de mapas de uso e ocupação de solo, produzidos através da classificação automática de imagens satélite. A partir desses mapas, é avaliada a extensão da ocupação de classes alvo, tais como áreas artificializadas (e.g. vias de comunicação, áreas residenciais, zonas industriais, zonas de extracção de minério, etc.), zonas ardidas, solo nu, etc. Uma elevada ocupação dessas classes, em detrimento de outras, como florestas e zonas húmidas, revela uma área com indícios de desertificação (Panigada et al., 2009). Deste modo, os métodos de classificação automática de imagens de satélite representam uma ferramenta fundamental para o desenvolvimento de políticas de protecção dos recursos naturais contra o fenómeno da desertificação.

O presente trabalho foi desenvolvido no âmbito do projecto DesertWatch-Extention (DWE), que vem no seguimento do DesertWatch-Original (DWO) e é financiado pela Agência Europeia Espacial (ESA –

European Spacial Agency), sendo os parceiros no Consórcio do projecto o Instituto Geográfico Português (IGP), a empresa Critical Software, o Instituto Superior Técnico - Centro de Recursos Naturais e Ambiente (IST - CERENA) e a empresa DEIMOS Engenharia. O DWE tem como objectivo principal o desenvolvimento e produção de mapas de indicadores de desertificação a partir de imagens de satélite, através de uma aplicação informática a ser desenvolvida no decurso do projecto, e o melhoramento de alguns aspectos do DWO. O DWE incide sobre três países (Portugal, Brasil e Moçambique) e pretende que os indicadores de desertificação sejam utilizados na monitorização deste fenómeno, de modo a auxiliar o cumprimento das suas obrigações na prestação de informação à Convenção das Nações Unidas de Combate à Desertificação (*United Nations Convention to Combat Desertification* – UNCCD).

A UNCCD procura garantir a produtividade a longo termo das zonas áridas desabitadas, o que não foi bem sucedido nos esforços anteriores. Em 2008, cerca de 192 governos assinaram a UNCCD, de modo a motivar a criação de parcerias e o desenvolvimento de directivas sobre programas de combate à desertificação. Em particular, o comité para a Ciência e Tecnologia da UNCCD (*Commitee on Science and Technology* - CST), desenvolveu um plano de directivas estratégicas projectadas para um período de 10 anos que visa a concretização dos objectivos proposto na UNCCD. Em particular, o plano desenvolvido pela CST propõe a monitorização e avaliação de factores biofísicos e socioeconómicos associados à desertificação e à degradação do solo de modo a suportar sistemas de apoio à decisão sobre políticas de gestão da água e do solo. O DWE foi desenvolvido tendo particular atenção a esta última directiva.

O projecto DWE encontra-se ainda ao abrigo do programa *Data User Element* (DUE) da ESA. O DUE tem como objectivo criar e reforçar as relações entre os utilizadores de dados espaciais de observação da Terra, sendo a disponibilização de aplicações informáticas

(preferencialmente de domínio publico) alimentadas com dados de observação da Terra um dos meios para concretizar esse objectivo, que respondam às necessidades dos utilizadores (e.g. mapas de indicadores de desertificação).

Assim, o objectivo do IGP no projecto DWE é rever e melhorar a nomenclatura de classe de uso/ocupação de solo e a metodologia de classificação utilizada no DWO, onde o mapa final deverá possuir uma precisão temática global de, pelo menos, 80%. Para mais, considerando ainda as linhas de orientação da DUE, procurou-se a simplificar a metodologia de classificação de modo a facilitar o papel do utilizador final da aplicação informática.

1.2 Revisão do Estado da Arte

O processo de classificação consiste na atribuição de *labels* a objectos (Fukunaga, 1991). Os objectos são entidades matemáticas compostas por um conjunto de atributos, usualmente numéricos. Esses objectos são abstracções de entidades do mundo real. Segundo Kuncheva (2004), o processo de classificação compreende os seguintes passos: i) definição do problema; ii) definição dos atributos e recolha da informação; e iii) classificação. No caso particular da classificação de imagens de satélite para a produção de mapas de uso e ocupação do solo (*Land Use / Land Cover* - LULC), poder-se-á ainda incluir um quarto passo: iv) validação do mapa resultante da classificação. No primeiro passo do processo de classificação procura-se definir os tipos de entidades do mundo real que se pretende classificar, i.e. o conjunto de classes a serem identificadas, a **nomenclatura**.

No segundo passo, procura-se definir o conjunto de atributos a serem utilizados pelo algoritmo de classificação e a fonte de dados. No caso da classificação de imagens satélite, a fonte de dados pode ser entendida como sendo as imagens de satélite a serem utilizadas e o conjunto de atributos as diferentes bandas espectrais e sintéticas (e.g. N*ormalized*

Difference Vegetation Index - NDVI). A selecção de um conjunto de dados adequado ao problema é um passo crítico que irá condicionar em grande medida o processo de classificação (Lu e Weng, 2004). De facto, devido às semelhanças espectrais entre classes LULC, um determinado conjunto de atributos pode não ser suficientemente informativo para discriminar correctamente algumas das classes; e, por outro lado, um conjunto suficientemente grande de atributos pode implicar perda da exactidão global (Hughes, 1968). Assim, é fundamental seleccionar, dentro dos dados disponíveis, apenas os atributos que garantam a máxima separabilidade entre classes LULC (Lu e Weng, 2004).

O terceiro passo do processo de classificação (a classificação propriamente dita) pode ser dividido em dois tipos: classificação supervisionada e classificação não supervisionada (Kuncheva, 2004). Este último caracteriza-se pela selecção de um processo automático de detecção de agrupamentos (*clusters*) de objectos semelhantes entre si, segundo um determinado critério de semelhança definido previamente. O resultado deste passo é uma classificação sem correspondência directa com as classes da nomenclatura. Posteriormente, um operador irá criar essa correspondência, classificando cada um desses agrupamentos.

A classificação supervisionada[1], por outro lado, requer a definição de uma amostra de treino antes da aplicação do algoritmo de classificação, i.e. de um conjunto de objectos classificados por um operador, de modo a serem utilizados como amostras das classes da nomenclatura. A introdução de conhecimento *a priori* dos objectos a serem classificados no algoritmo de classificação é vantajosa, uma vez que estes classificadores tendem a produzir resultados mais exactos que os algoritmos não supervisionados (Manter, 2004). Porém, a qualidade do resultado deste tipo de classificadores está fortemente dependente da qualidade da amostra de treino (Mather, 2004). Esta depende essencialmente de dois factores: dimensão e representatividade (Chen e

[1]No presente texto, quando se falar de classificação, estará subentendido que a classificação é supervisionada.

Stow, 2002). De um modo geral, procura-se que a amostra de treino seja suficientemente grande para permitir o cálculo de uma estimativa robusta dos parâmetros estatísticos das classes da nomenclatura (fundamental em classificadores de natureza estatística, como é caso do classificador de máxima verosimilhança) e representativa o bastante de modo a incluir toda a variabilidade da população de objectos a serem classificados (Mather, 2004). Contudo, recolher uma amostra "grande" é um processo que tende a consumir bastante tempo e, apesar desta recolha usualmente ser feita sobre níveis de informação de resolução mais detalhada (e.g. fotografias aéreas, mapas anteriormente produzidos e imagens satélite), a introdução de erros devido à natureza subjectiva da interpretação de fotografias aéreas e imagens de satélite é quase inevitável (Cigolani et al, 2004).

Wilkinson (2005) apresenta uma revisão bibliográfica dos últimos 15 anos de experiências com classificação de imagens de satélite e conclui que, apesar do esforço em introduzir métodos mais sofisticados de classificação de imagens (e.g. redes neuronais), assim como métodos de recolha de amostras de treino e de análise de separabilidade de classes, não existiu uma melhoria considerável na qualidade dos produtos. Wilkinson aponta que, durante este período, o desenvolvimento foi realizado essencialmente em três vertentes: primeira, foram apuradas as componentes da classificação, como a recolha de amostras de treino, métodos de análise de separabilidade entre classes e a criação de novos índices de separabilidade; segunda, o desenvolvimento de novas abordagens que procuram melhorar os classificadores já existentes, como a introdução de classificadores fuzzy e a combinação de classificadores; e terceira, a fusão de diferentes tipos de sensores e fontes de informação. Segundo Wilkinson existem três possíveis explicações: primeira, talvez os desenvolvimentos mais promissores nas técnicas de classificação não tenham sido devidamente assimilados pela comunidade de investigadores de detecção remota; segunda, talvez a velocidade do desenvolvimento de novas técnicas de exploração de imagens de satélite não consiga acompanhar o aumento da

complexidade da informação associada aos sensores de observação da terra; e terceira, talvez os desenvolvimentos científicos sejam incompletos na incorporação da subjectividade humana no produto da classificação. Deste modo, o estudo de Wilkinson sugere que os desenvolvimentos futuros não passam pelo refinamento dos algoritmos de classificação, mas antes pelo estudo da definição das classes, da relação entre escala e classificação e pela introdução de medidas de subjectividade na avaliação da qualidade da classificação.

Os estudos sobre a aplicabilidade dos classificadores compostos na classificação de imagens de satélite existem desde que 1994 (Conese e Maselli, 1994). Esses estudos têm-se concentrado na combinação de classificadores assistidos com classificadores não assistidos (Lo e Choi, 2004), na combinação de classificadores "simples", como o classificador de máxima verosimilhança, com classificadores mais sofisticados, como as redes neuronais e árvores de classificação (Warrender e Augusteihn, 1999; Lu e Weng, 2004), e na exploração de regras de classificação compostas, como a regra do produto[2] (Steele, 2000). De um modo geral, os estudos mostram que existe um ganho na exactidão temática do mapa final, especialmente em imagens de satélite de elevada resolução (Kuncheva, 2004). Contudo, esses estudos têm-se limitado à aplicação de estratégias de sucesso noutras áreas da aplicação da classificação, como no caso da do reconhecimento de caracteres, e introduzido pouca informação sobre as classes de uso e ocupação de solo no processo de classificação.

Wilkinson (2005) sugere que as estratégias para a classificação automática de imagens de satélite não introduziram melhorias significativas. Entre essas estratégias inclui-se a abordagem do multiclassificador (ou, por outro nome, classificador composto). A aplicação deste tipo de classificadores no processo de classificação de

[2]A regra do produto consiste numa generalização possível da regra do máximo. Na regra do máximo procura-se a classe que maximize a probabilidade *a posteriori* segundo um determinado classificador. Na regra do produto procura-se a classe que maximize o produto de probabilidades *a posteriori* associadas a diferentes classificadores.

imagens de satélite tem revelado que as estratégias sugeridas são conservadoras no que toca à sua aplicabilidade no caso específico da classificação de imagens de satélite, com a finalidade de produzir mapa LULC. Assim, no presente estudo, procura-se explorar um modo de introduzir a informação sobre a disposição das classes no espaço de classificação, em particular a proximidade informacional entre classes espectrais. Para isso, pretende-se melhorar os resultados obtidos com os classificadores simples, por meio da composição de classificadores e de uma medida de entropia informativa de modo a restringir a lista de possíveis classes a serem atribuídas ao *pixel* a ser classificado.

1.3 Organização da Dissertação

O presente documento é composto por quatro partes. A primeira, secção 2, apresenta um resumo dos conceitos e dos métodos de classificação automática. Procurou-se nesta secção apresentar os elementos mais genéricos e fundamentais no processo de classificação, independentemente a área de aplicação. A segunda parte, secção 3, apresenta os dados utilizados e a área em que foi realizado o estudo, assim como os passos mais importantes da metodologia aplicada. Na terceira parte, secção 4, são apresentados os resultados obtidos no presente estudo. Na última parte, secção 5, é apresentada a conclusão do estudo.

2 Métodos de Classificação Automática

Nesta secção é apresentada a terminologia fundamental na classificação e o modelo mais genérico de um classificador; seguem-se os elementos probabilísticos do processo de classificação e uma descrição de cada um dos classificadores singulares utilizados neste estudo. No final são apresentados os conceitos elementares dos classificadores compostos.

2.1 Conceitos Preliminares

Os conceitos preliminares ao processo de classificação são, essencialmente, seis: espaço de classificação, classe, atributo, conjunto de dados, treino e teste. O **espaço de classificação** pode ser entendido como o conjunto de todos os objectos a serem classificados; a **classe**, um subconjunto de objectos do espaço de classificação, de tal modo definida que os objectos dessa classe não pertencem a mais nenhuma classe. O **atributo** representa uma característica do objecto que, de um modo geral, pode ser definida como sendo uma quantidade escalar. O termo **conjunto de dados** encontra-se associado ao conjunto de objectos que entra no processo de classificação, usualmente estruturado em matriz, onde cada linha define um objecto do espaço de classificação e onde a primeira coluna possui o *label* da classe a que pertence o objecto; as restantes colunas são os atributos consideradas na classificação. O **treino** e o **teste** são conjuntos de dados*:* o primeiro utilizado para treinar o classificador e o segundo utilizado para avaliar a qualidade do classificador. Kuncheva (2004) e Hastie et al. (2009) definem ainda um terceiro tipo de conjunto de dados, a **validação**, que serve para afinar o algoritmo de classificação caso este possua parâmetros de entrada, dos quais se pretende determinar os valores óptimos. No presente caso, este tipo de *conjunto de dados* não foi definido. Assim, neste estudo, o espaço de classificação é o conjunto de *pixels* da imagem a ser classificada. Cada *pixel* é um vector com tantas componentes quanto o número de bandas da imagem. Cada componente do *pixel* é um atributo no processo de classificação.

2.2 Modelo Canónico

Seja $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ o conjunto dos *labels* de todas as classes do problema, onde c é o número de classes consideradas. Então, um classificador é definido como sendo qualquer função D tal que $D{:}E \subset \mathbb{R}^n \rightarrow \Omega$, onde n é o número de dimensões do problema e E é o espaço de classificação (i.e. o conjunto dos objectos a serem classficados). Esta é a definição mais genérica de um classificador. Contudo, para que se

torne um conceito mais prático, é necessário explicitar a regra de classificação. O **modelo canónico da classificação** (Kuncheva, 2004), ilustrado na figura 1, define essa regra.

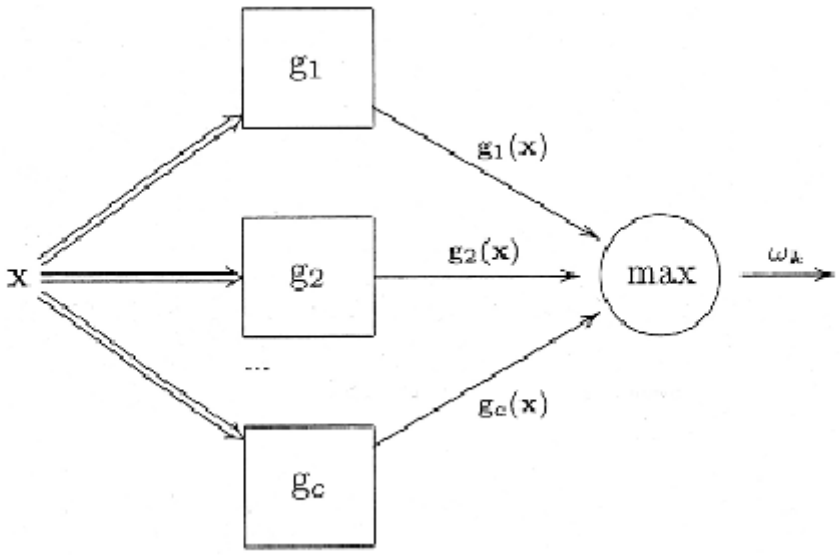

Figura 1 – Modelo canónico da classificação (adaptado de Kuncheva, 2004).

Este modelo impõe a existência das funções $g_i: \mathbb{R}^n \rightarrow \mathbb{R}$, onde i = 1, …, c, denominadas **funções discriminantes**, que atribuem uma pontuação a cada classe do problema, para um qualquer objecto do espaço de classificação x. Nestas condições, a regra de classificação mais imediata é definida pela **regra do máximo** (Kuncheva, 2004), i.e. para qualquer objecto$x \in E$do espaço de classificação, o *label* a ser atribuído a x, i.e. D(x), é o *label* da classe j tal que $g_j(x) = max\{g_1(x), g_2(x), \ldots, g_c(x)\}$[3].

As funções discriminantes irão particionar o espaço de classificação E em c regiões diferentes, denominadas por **regiões de decisão** (Figura 2). Cada uma delas é definida pela regra do máximo do seguinte modo: R_i é o espaço geométrico em $\mathbb{R}^n$ em que a função discriminante

[3]Com esta regra é possível que diferentes classes tenham a mesma pontuação e, portanto, é possível haver dois ou mais máximos. Assim, para completar o modelo, é necessário definir uma regra de desempate. Esta, usualmente, pode ser definida pela escolha aleatória das classes empatadas (Kuncheva, 2004).

associada à classe ω_i é máxima, i.e.:

$$R_i = \{x \in E : \omega_i = D(x)\} \quad (1)$$

para i = 1, 2, …, c; consequentemente, todos os objectos do espaço de classificação contidos em$R_i$irão receber o *label* ω_i. Às fronteiras entre regiões chamar-se-ão **fronteiras de decisão** (Figura 2). Os objectos que se encontram na fronteira são os objectos que possuem mais que uma classe com pontuação máximo e, portanto, são objectos onde é necessário aplicar a regra de desempate para definir o seu *label* final.

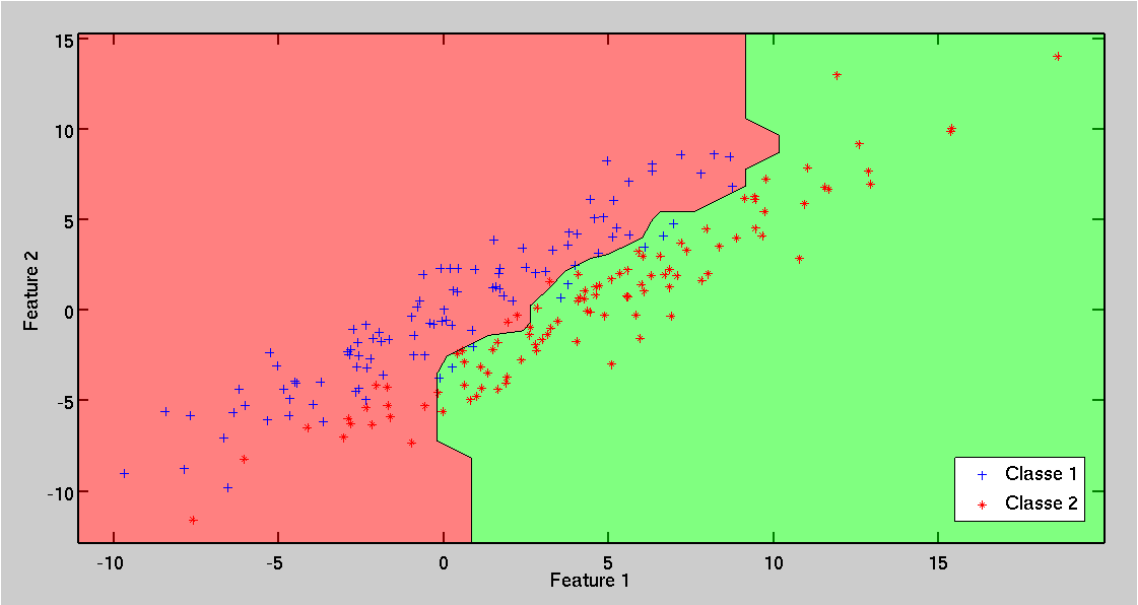

Figura 2 – Regiões de decisão e a fronteira de decisão gerados com o classificador *3-neireast neighbor* com dados sintéticos 2D. A rosa a região de decisão do espaço de classificação associado à classe 1 e a verde a região associada à classe 2.

2.3 Elementos Probabilísticos da Decisão

Os conceitos introduzidos anteriormente não apresentam ainda os elementos necessários para se tornarem aplicáveis a situações reais.

Como o processo de classificação é, por natureza, um processo não-determinístico, é natural introduzir conceitos probabilísticos na regra de decisão. Assim, se admitirmos que o *label* a ser atribuído a qualquer$x \in E$é uma variável aleatória discreta que toma valores do conjunto $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$, então podemos definir a **probabilidade *a priori*** da classe$\omega_i$para qualquer i = 1, ..., c, como sendo o valor da função massa de probabilidade $P(\omega_i)$. A função massa de probabilidade raramente é conhecida, pelo que na prática, esta pode ser estimada pela frequência da classe na amostra de treino ou impondo uma distribuição uniforme caso não se pretenda favorecer à partida nenhuma das classes (Hastie et al, 2009).

Suponhamos agora que os objectos da classe$\omega_i$se encontram distribuidos no espaço de classificação E segundo um modelo estocástico$P(x \mid \omega_i)$. Deste modo, é possivel, por meio da Regra de Bayes, inferir a **probabilidade *a posteriori*** associada ao par (ω_i, x), i.e.

$$P(\omega_i \mid x) = \frac{P(\omega_i)P(x|\omega_i)}{\sum_{j=1}^{c} P(\omega_j)P(x|\omega_j)} \quad (2)$$

A probabilidade *a posteriori* pode ser, neste caso, entendida como a resposta à pergunta "dado o objecto$x \in E$, qual a probabilidade de x pertencer à classeω_i?" Deste modo, é imediato que uma possível regra de classificação consiste na atribuição do label da classe que maximize a probabilidade *a posteriori*. Portanto, as funções de probabilidade *a posteriori* tornam-se assim possíveis funções discriminantes, pelo que podemos rescrever a regra do máximo da seguinte forma:

$$D(x) = \omega_i \in \Omega,\ se \omega_i = argmax_{\omega_j}\{P(\omega_j \mid x)\} \quad (3)$$

Ou seja, o *label* a atribuir ao objecto $x \in E$ é o *label* da classe que possui o valor de probabilidade *a posteriori* máxima associada ao objecto x.

Assim, as funções discriminantes podem ser analiticamente expressas por $g_i(x) = P(\omega_i)P(x \mid \omega_i)$, que constitui apenas o numerador da regra de Bayes, uma vez que o denominador é comum para todas as classes, pelo que não altera a ordenação das classes segundo a probabilidade *a posteriori*. Uma alternativa frequente para a definição das funções discriminantes é feita por meio do logaritmo, $g_i(x) = \log(P(\omega_i)P(x \mid \omega_i))$, que também não altera a ordenação das classes pela monotonia da função logaritmica. Aos classificadores que baseiam o seu processo na maximização da probabilidade *a posteriori* chamam-se **classificadores bayesianos**.

2.4 Classificadores Singulares

Nesta secção apresentam-se os classificadores singulares comparados neste estudo. Os classificadores foram seleccionados tendo em conta duas condições: a primeira é que deveriam ser classificadores *state-of-the-art*; a segunda condição é que não necessitassem de parametrização ou, caso tivessem parâmetros de entrada, a sua optimização fosse simples, de modo a minimizar e simplificar a interacção com o utilizador final do *software*. Assim, os classificadores seleccionados foram: o classificador linear discriminante, o classificador quadrático discriminante, o classificador de Parzen, o classificador *k-nearest neighbor* e as árvores de classificação.

2.4.1 Classificadores Paramétricos

No presente texto, o termo "classificador paramétrico" encontra-se associado aos classificadores que partem da premissa de que as classes dos objectos do espaço de classificação são uma população gaussiana. Nestas condições, os classificadores paramétricos podem ser divididos em classificadores lineares e classificadores quadráticos.

Os classificadores linear e quadrático recebem o seu nome do tipo de função discriminante que utilizam no processo de classificação. De um

modo geral, se um classificador define hiperplanos de separabilidade, dir-se-á que é um **classificador linear**; por outro lado, se a função discriminante é uma forma quadrática, dir-se-á que é um **classificador quadrático**.

Para construir este tipo de classificadores poder-se-á recorrer à regra do máximo e impor que as funções discriminantes destes classificadores são tais que, para qualquer $x \in E$ e i=1, …, c,

$$g_i(x) = \log(P(\omega_i)P(x \mid \omega_i)) \quad (4)$$

Por hipótese, todas as classes têm um comportamento normal, assim suponhamos μ_i e Σ_i para a classeω_i, pelo que a equação 1, pode ser reescrita na forma (Kuncheva, 2004):

$$g_i(x) = \log P(\omega_i) - \frac{n}{2} log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) \quad (5)$$

Se, agora, for imposta a condição de homocedasticidade, i.e. as classes possuem igual variância (para qualquer i, $\Sigma = \Sigma_i$), então a equação 2 fica reduzida a uma forma linear com expressão:

$$g_i(x) = w_{i0} + w_i^T x \quad (6)$$

onde $w_i = \Sigma^{-1}\mu_i$ e $w_{i0} = \log P(\omega_i) - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i$ define o termo independente dos hiperplanos de separabilidade. Nestas condições, a variância comum pode ser definida por meio da média simples das variâncias das classes ou pela média ponderada pela probabilidade *a posteriori* de cada classe (Kuncheva, 2004). Fica assim definido o classificador discriminante linear (*Linear Discriminant Classifier* - LDC).

Por outro lado, se a condição de homocedasticidade não for imposta, a equação 2 pode ser manipulada, assumindo a expressão (Kuncheva, 2004):

$$g_i(x) = w_{i0} + w_i^T x + x^T W_i x \quad (7)$$

onde $w_{i0} = \log P(\omega_i) - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i - \frac{1}{2}\log|\Sigma_i|$ e $W_i = -\frac{1}{2}\Sigma_i^{-1}$, definindo deste modo o classificador discriminante quadrático (*Quadratic Discriminant Classifier* – QDC), conhecido também por classificador de máxima verosimilhança (Mather, 2004).

O procedimento para a classificação com qualquer um destes classificadores está descrita no algoritmo 1 (Quadro 1).

Algoritmo 1 – Classificador LDC / QDC

Input:

i) Amostra de treino para cada uma das classes;

ii) Regra de desempate.

Output:

Espaço de classificação E classificado.

Procedimento:

1. Para todas as classes da amostra de treino:

1. 1. Calcular o vector valor-médio;

1. 2. Calcular a matriz de covariâncias;

2. Se o classificador é o LDC, calcular a matriz de covariância média;

3. Para todo o objecto $x \in E$:

3. 1. Para cada classe do problema, calcular o valor da função $g_i(x)$;

3. 2. Escolher a(s) classe(s) com valor discriminante máximo; 3. 3. Se existirem classes empatadas, aplicar regra de desempate; 3. 4. Atribuir o *label* da classe resultante a x.

Quadro 1 – Algoritmo do LDC / QDC.

Estes dois classificadores são conceptualmente simples e, em termos computacionais, eficientes tanto em tempo de execução como em uso de memória; e apesar das suas condições de partida serem restritas, quando comparadas com as dos restantes classificadores, a prática mostra que são suficientemente precisos e robustos mesmo quando as circunstâncias do problema se afastam das ideais (Kuncheva, 2004; Hastie et al., 2009).

2.4.2 Classificadores Não Paramétricos

Os classificadores anteriores imponham que as classes do problema possuíssem uma estrutura modelada pela distribuição normal. Contudo, essa condição torna o LDC e o QDC teoricamente pouco apropriados em problemas nos quais se desconhece o comportamento estatístico dos dados.

Os próximos classificadores (classificador de Parzen e k-NN) permitem ultrapassar a necessidade de se impor a condição da normalidade nos dados, estimando a probabilidade $P(x \mid \omega_i)$ numa vizinhança do objecto x a ser classificado. Seja q a probabilidade do objecto$x \in E$se encontrar no subespaço $R \subset E$ e $p = P(x \mid \omega_i)$ a distribuição desconhecida dos dados, resulta então que (Pestana e Velosa, 2002):

$$q = P(x \in R) = \int_R p(u)du \quad (8)$$

Suponhamos que é realizada uma amostragem de N elementos da classe

ω_i do treino. Então, a probabilidade de exactamente k elementos desses N se encontrarem no subespaço R é modelada pela distribuição binomial com parâmetros (q, N), pelo que um estimador para q será dado por:

$$\hat{q} \approx \frac{k}{N} \quad (9)$$

Escolhendo uma região R suficientemente pequena, a continuidade da função de distribuição implica que, em R, q é aproximadamente uniforme, pelo que

$$\hat{q} \approx p(x) \int_R du = p(x) V_R \quad (10)$$

onde V_Ré o hiper-volume da região R. Recorrendo à equação 5, ficamos com $\frac{k}{N} \approx p(x) V_R$, e como $p = P(x \mid \omega_i)$, conclui-se que:

$$P(x \mid \omega_i) \approx \frac{k}{N V_R} \quad (11)$$

A equação 6 é o ponto de partida para os classificadores não-paramétricos (Kuncheva, 2004; Fukunaga, 1991). Duas características comuns aos classificadores paramétricos é a sua complexidade computacional, ou seja, estes classificadores tendem a utilizar muita memória e são usualmente lentos, e o facto de sofrerem do fenómeno *curse of dimensionality*[4] (Kincheva, 2004; Hastie et al, 2009). Este fenómeno consiste na relação que existe entre a dimensionalidade do

[4]O fenómeno curse of dimensionality, ou fenómeno de Hughes, é um fenómeno que se verifica em problemas de classificação e que consiste na relação entre a complexidade do problema e a exactidão global. De um modo geral, este fenómeno é caracterizado pela redução da exactidão global da classificação à medida que a complexidade do problema aumenta. A complexidade de um problema de classificação é usualmente determinado pelo número de dimensões do espaço de classificação (Hughes, 1967).

problema e o número de indivíduos de treino necessários para um classificador não-paramétrico realizar a sua tarefa; de um modo geral, este tende a aumentar de modo bastante rápido à medida que o número de dimensões aumenta (Hastie et ali, 2009).

2.4.2.1 Classificador de Prazen

Neste classificador, recorremos à equação 11 e fixamos as variáveis N eV_R, e estimamos k a partir dos dados. A forma mais intuitiva para estimar k consiste em percorrer a amostra de treino para uma determinada classe e contar o número de elementos que se encontram na vizinhança, previamente definida, do objecto a ser classificado. Contudo, essa abordagem implica que a amostragem tenha sido realizada sobre uma população com distribuição uniforme, o que raramente é o caso (Web, 2002). Deste modo, é necessário um método mais robusto de se estimar a distribuição estatística dos dados na vizinhança de um objecto; esse método é o *kernel estimator.*

Uma função *kernel estimator* é uma função K tal que:

$$\int_{\mathbb{R}^n} \frac{1}{h^n} K(\frac{x-z_j}{h}) dx = 1 \quad (12)$$

para qualquer j de 1 até ao número total de elementos da amostra. O parâmetro h é o chamado parâmetro de suavização do estimador. Para o caso multidimensional, os estimadores mais usuais são o estimador uniforme e o estimador gaussiano, sendo este último o mais frequente (Fukunaga, 1991). A expressão analítica para o estimador gaussiano é (Kuncheva, 2004):

$$\frac{1}{h^n} K_G\left(\frac{x-z_j}{h}\right) = \frac{1}{h^n (2\pi)^{\frac{n}{2}} \sqrt{|S|}} \exp(\frac{-d^2(x,z_j)}{2h^2}) \quad (13)$$

Onde d^2 é o quadrado da distância de Mahalanobis e S uma matriz simétrica (semelhante à matriz de covariâncias) que define a forma do estimador. Assim, substituído a função estimadora na equação 6, ficamos com:

$$P(x \mid \omega_i) = \frac{1}{N_i} \sum_{z \in \omega_i} \frac{1}{h^n} K_G\left(\frac{x-z}{h}\right) \quad (14)$$

Onde N_i é o número de elementos da amostra de treino da classe ω_i. Prova-se (Kuncheva, 2004) que a probabilidade *a posteriori* se relaciona com a função kernel pela expressão 7

$$P(\omega_i \mid x) \propto \sum_{z \in \omega_i} K_G\left(\frac{x-z}{h}\right) \quad (15)$$

Assim, se a regra do máximo for aplicada, pode-se definir a função discriminante para o classificador de Parzen com estimador gaussiano pela equação 8:

$$g_i(x) = \sum_{z \in \omega_i} K_G\left(\frac{x-z}{h}\right) \quad (16)$$

Segue-se o algoritmo de classificação com o classificador de Parzen com kernel gaussiano (Quadro 2).

Algoritmo 2 – Classificador de Parzen com kernel gaussiano.

Input:

i) Parâmetro de suavização h;

ii) Matriz de forma S;

iii) Amostra de treino;

iv) Espaço de classificação E;

v) Regra de desempate.

Output:

Espaço de classificação E classificado.

Procedimento:

1. Calcular a raiz quadrada do determinante da matriz de forma S;

2. Para cada objecto $x \in E$:

2. 1. Para cada elemento z do treino, calcular$K_G(\frac{x-z}{h})$;

2. 2. Para cada classe do problema ω_i, calcular o valor da função $g_i(x)$;

2. 3. Determinar a(s) classe(s) com valor discriminativo máximo;

2. 4. Se existirem classes empatadas, aplicar regra de desempate;

2. 5. Aplicar o *label* da classe resultante a x.

Quadro 2 – Algoritmo do classificador de Parzen.

O classificador de Parzen apresenta condições teoricamente mais genéricas que o LDC e o QDC. Contudo, a sua aplicação em problemas práticos não é tarefa trivial, uma vez que o parâmetro de suavização está fortemente relacionado os dados do problema (Webb, 2002). Para mais, o classificador de Parzen encaixa-se na categoria dos classificadores *memory-based*, i.e. para cada objecto que se pretenda classificar, é necessário ter acesso a cada um dos elementos da amostra de treino, o que implica uma elevada exigência de espaço memória para a computação. Finalmente, o classificador de Parzen necessita de calcular a distância de Mahalanobis do objecto-alvo para cada um dos elementos do treino, o que torna o algoritmo consideravelmente mais

lento que os classificadores paramétricos (Hastie et al, 2009).

2.4.2.2 Classificador de k-NN

Anteriormente no classificador de Parzen, recorreu-se à equação 6, na qual se fixaram as variáveis V_R e N. Para o classificador k-NN, iremos fixar as variáveis k e N. Dado que a probabilidade *a posteriori* de um objecto genérico x pertencer à classe ω_i é dada por:

$$P(x \mid \omega_i) \approx \frac{k_i}{N_i V_r} \quad (17)$$

onde k_ié o número de objectos da amostra da classe ω_i, N_i o número de elementos na amostra e V_R o hiper-volume da região R. Para a definição da região R, poder-se-á recorrer a qualquer métrica no espaço de classificação (Kuncheva, 2004), sendo as mais usuais a distância euclidiana (nesse caso R será uma hiper-esfera) e a distância de *Manhattan* (neste caso R será um hiper-cubo). Assim, aplicando-se a regra de *Bayes*, temos que:

$$P(\omega_i \mid x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} \approx \frac{\frac{k_i}{N_i V_R}\frac{N_i}{N}}{\frac{k}{NV_R}} \quad (18)$$

e portanto $P(\omega_i \mid x) \approx \frac{k_i}{k}$, ou seja pode-se colocar$g_i(x) = \frac{k_i}{k}$. Assim, pela regra do máximo, a classe que maximiza a probabilidade *a posteriori* é a classe mais frequente nos k vizinhos mais próximos. Segue-se o algoritmo de classificação k-NN (Quadro 3):

Algoritmo 3 – Classificador k-NN

Input:

i) Número de vizinhos k;

ii) Métrica a ser utilizada d;

iii) Regra de desempate;

iv) Amostra de treino;

v) Espaço de classificação E.

Output:

Espaço de classificação E classificado.

Procedimento:

1. Para cada objecto $x \in E$:

1. 1. Para cada elemento z do treino, calcular d(x, z);

1. 2. Ordenar as distâncias calculadas descendentemente;

1. 3. Escolher as k primeiras distâncias;

1. 4. Determinar a(s) classe(s) mais frequente(s) das k distâncias;

1. 5. Se existirem classes empatadas, aplicar regra de desempate;

1. 6. Atribuir o *label* da classe resultante a x.

Quadro 3 – Algoritmo do classificador k-NN.

O classificador k-NN tem a vantagem de ser conceptualmente simples e de ser um classificador exclusivamente geométrico, i.e. não se baseia em distribuições estatísticas, que são usualmente desconhecidas. Contudo, tal como o classificador de Parzen, possui um parâmetro que requer afinação e que está fortemente relacionado com a estrutura dos dados de entrada, o que torna a sua optimização difícil (Fukunaga, 1991; Kuncheva, 2004; Hastie et al., 2009). Tal como o classificador Parzen, o k-NN necessita de visitar todos os elementos da amostra de treino para cada classificação que realiza, o que o torna computacionalmente exigente. Para mais, o k-NN é bastante sensível à

escolha da métrica, à escolha da amostra de treino e aos *outliers* (Kuncheva, 2004). Finalmente, o k-NN actua como uma caixa-preta sobre os dados, o que significa que fornece pouca informação relativamente à sua estrutura (Kuncheva, 2004; Hastie et al., 2009).

2.4.3 Árvores de Classificação

As árvores de classificação são do mesmo tipo de classificadores considerados até ao momento, i.e. classificadores bayesianos (Kuncheva, 2004). Assim, o resultado das árvores de classificação não garante que o *label* atribuído a um determinado objecto seja aquele que maximiza a probabilidade *a posteriori*. As árvores de classificação são conhecidas pela sua capacidade de definir fronteiras de decisão extremamente complexas, o que leva estes classificadores a ficarem demasiadamente ajustados aos dados de treino com facilidade, fenómeno usualmente designado por *overfitting* (Kuncheva, 2004). De facto, para qualquer amostra de treino, as árvores de classificação podem produzir um resultado em *cross-validation* de 100%, desde que ela possa crescer sem qualquer restrição. O processo de classificação nas árvores de classificação pode ser visto como uma sequência de decisões simples que culminam no *label* a ser atribuído ao objecto. A construção de uma árvore de classificação inicia-se no topo, i.e. nodo inicial ou raiz. Seguidamente, o nodo inicial é dividido em nodos-filho; estes por sua vez são divididos em subnodos e assim sucessivamente até se alcançar o critério de paragem. Assim, no algoritmo para a construção de uma árvore de classificação existem três processos essenciais: a divisão dos nodos e o critério de paragem do processo de divisão.

O modo mais usual de definir a divisão dos nodos é por meio da divisão binária (Kuncheva, 2004). Na divisão binária, cada nodo tem exactamente dois filhos, onde por exemplo o nodo da esquerda corresponde à resposta negativa e o nodo da direita à resposta positiva à questão associada ao nodo. Essa questão será geralmente da forma

"$x \leq x_0$?" onde x é uma *atributo* e x_0 é um valor de *threshold*. O problema que se coloca é: que *atributo* e valor de *threshold* utilizar na questão do nodo?

Suponhamos que $P_j(t)$ é a probabilidade de pelo menos um elemento da classe j chegar ao nodo t. Esta probabilidade pode ser estimada pela proporção de elementos da classe j que chegam até ao nodo t. A impureza de um nodo t é a distribuição dos *labels* das classes em t (Webb 2002; Kuncheva, 2004). Intuitivamente, a impureza mede a variabilidade existente num determinado nodo: quanto mais variabilidade existir num nodo, menos puro será o nodo. Existem diversas medidas para a medição da pureza de um nodo, por exemplo, a impureza baseada no índice de Gini, a impureza baseada na entropia e a impureza baseada nos erros de classificação. A medida mais frequentemente utilizada é a impureza baseada no índice de Gini (Webb, 2002; Kuncheva, 2004). O procedimento para a determinação da *atributo* e do valor de *threshold* consiste na selecção da *atributo* e do valor de *threshold* que miminiza a impureza do nodo. Esse processo de selecção usualmente é realizado por meio de um algoritmo ganancioso[5], o que não garante que a árvore de classificação seja óptima, segundo qualquer critério.

O critério de paragem da divisão é importante para garantir que a árvore de classificação não caia em *overfitting* (Webb, 2002; Kuncheva, 2004). A questão é: em que condições devemos parar o processo de divisão? Esta pergunta não tem uma resposta definitiva (Webb, 2002). De facto, o modo mais frequente consiste em parar o processo de divisão quando o nível de pureza chegar a um determinado valor fixo à partida (Webb, 2002; Kuncheva, 2004). Contudo, esse valor é de difícil afinação (Webb, 2002), uma vez que se for demasiadamente elevado, a árvore de classificação recebe pouco treino (fenómeno designado por *undertraining* ou, alternativamente, *underfitting*); mas se for

[5]Um algoritmo ganancioso é um tipo de algoritmo de busca de soluções a um determinado problema que em cada passo escolhe a melhor solução no momento, segundo um determinado critério.

demasiadamente baixo, a árvore tenderá a cair em *overfitting*. No algoritmo utilizado no presente trabalho, o critério de paragem é definido por meio de um teste de hipóteses que avalia se a próxima divisão é benéfica ou não para a separabilidade das classes. Determinam-se duas quantidades, χ_L^2 e χ_R^2, associadas ao futuro nodo da esquerda e ao futuro nodo da direita, respectivamente. Estas quantidades são dadas por (Kuncheva, 2004):

$$\chi_L^2 = \sum_{i=1}^{c} \frac{(nn_{Li} - n_L n_i)^2}{nn_L n_i} \quad (19)$$

onde C é o número de classes, n é número total de elementos do treino que chegaram ao nodo-pai, n_i é o número de elementos da amostra de treino da classe i que chegaram ao nodo-pai, n_L é o número de elementos da amostra de treino que chegam ao nodo-filho da esquerda e n_{Li} é o número de elementos da amostra de treino da classe i que chegam ao nodo-filho da esquerda. A equação é analoga para a quantidade χ_R^2. Nestas condições, a média entre χ_L^2 e χ_L^2 é tem uma distribuição Qui-Quadrado com C-1 graus de liberdade. Assim, o critério de paragem é satisfeito quando a média de χ_L^2 e χ_L^2 for superior ao valor tabelado da função Qui-Quadrado com C-1 graus de liberdade (e para um determinado nível de significância de α). A alternativa será utilizar um valor de *threshold* imposto pelo utilizador e compará-lo com a média de χ_L^2 e χ_L^2. Os valores admissiveis para o threshold vão desde 0 a 10 (Kuncheva, 2004). Se o threshold for fixo no 0, então a paragem só ocorre quando já não existirem elementos do treino por classificar (o que leva ao *overfitting*). Se o valor for fixo em 10, isso irá tornar a árvore muito curta, o que pode colocar a árvore de classificação em condições de *underfitting* (Kuncheva, 2004). No presente trabalho optou-se por recorrer à possibilidade de impor um valor de *threshold*, uma vez que esta alternativa permite um maior controlo sobre o crescimento da árvore. Este algoritmo encontra-se implementado no

Matlab Pattern Recognition Toolbox (*PR toolbox*) disponibilizado pela Universidade Técnica de Delft, Holanda.

2.5 Classificadores Compostos

O termo "classificador composto" ou "multiclassificador" encontra-se associado a algoritmos de classificação que combinam diferentes tipos de classificadores simples (ou singulares) de modo a minimizar o erro de classificação. A filosofia por detrás destes sistemas é, então, maximizar a precisão da classificação sem aumentar de modo significativo a complexidade do algoritmo, compensando os pontos fracos de um classificador com os pontos fortes de outro (Kuncheva, 2004). Portanto, quando se fala em classificadores compostos não se pergunta qual o melhor classificador, mas antes qual a melhor combinação de classificadores para a tarefa (Hastie et al, 2009). Existem essencialmente dois tipos de classificadores compostos: os classificadores em **sistema competitivo** e os classificadores em **sistema cooperativo**.

No sistema competitivo cada classificador trabalha independentemente dos restantes, sem que um possa influenciar a decisão dos restantes (Kuncheva, 2004). O que significa que não há troca de informação entre algoritmos, e cada um apresenta a sua decisão de modo independente. A decisão final é resolvida por meio de uma regra externa aos classificadores. Existe uma multiplicidade de regras de resolução nos sistemas competitivos, sendo o voto maioritário e o sistema "em leilão" os mais comuns (Kuncheva, 2004). A regra do voto maioritário pode ser de três tipos: i) pluralidade (o *label* final é o *label* mais frequente); ii) maioria simples (o *label* a ser atribuído é a aquele que for atribuído por pelo menos 50% + 1 dos classificadores); e iii) unanimidade (o *label* a ser atribuído é aquele no qual todos os classificadores estão de acordo). No sistema "em leilão" o *label* a ser atribuído é dado pelo classificador que atribuir maior probabilidade *a posteriori* (Kuncheva, 2004).

No sistema cooperativo, por outro lado, os classificadores partilham informação no processo de classificação. Enquanto que no sistema competitivo o processo de decisão parece fragmentado, o que implica o uso de uma regra exterior aos classificadores para se chegar à decisão final, no sistema cooperativo, a interacção entre os classificadores constitui a própria regra de decisão. Os tipos mais comuns de sistema cooperativo são dois: o sistema baseado em áreas de competência e o sistema em cascata (Kuncheva, 2004). O sistema baseado em áreas de influência consiste em particionar o espaço de classificação em regiões associadas a classificadores particulares. Essas áreas são atribuídas em função da probabilidade *a posteriori* dos classificadores, de modo a fazer corresponder uma área do espaço de classificação ao classificador que atribua maior probabilidade *a posteriori* nessa região. O segundo tipo, o sistema em cascata ou sistema sequencial, consiste na classificação por meio de diversos varrimentos do espaço com diferentes classificadores. Por exemplo, um sistema em cascata poderia ser composto por uma sequencia de classificadores, A, B, C, ..., em que o procedimento seria: para cada objecto x do espaço de classificação, o classificador A classifica o objecto x e atribui probabilidade *a posteriori* p; se p estiver abaixo de um determinado *threshold*, então entra em acção o classificador B, e o processo repete-se até se chegar a uma classificação com uma probabilidade *a posteriori* superior ao *threshold* ou ao fim da sequência, e nesse caso uma regra de excepção deve ser aplicada (Kuncheva, 2004).

A aplicação dos classificadores compostos tem permitido aumentos de exactidão global comparativamente às abordagens singulares em diversas áreas de aplicação, tais como reconhecimento de caracteres orientais, no reconhecimento de rostos e na classificação de imagens de satélite, sem um acréscimo considerável na complexidade computacional.

3 Metodologia

Nesta secção são apresentados os principais passos da metodologia seguida para a classificação. São apresentadas a área de estudo, as imagens de satélite utilizadas na classificação, o procedimento seguindo para a selecção das amostras de treino, o protocolo para a validação dos mapas produzidos, o modo seguindo para a comparação dos classificadores e uma descrição genérica do classificador composto, onde são introduzidos os conceitos fundamentais sobre as medidas de entropia informativa aplicadas posteriormente.

3.1 Área de Estudo

A área de estudo para o presente estudo foi definida pelos utilizadores do sistema informático que o projecto DWE procurou desenvolver. Esta área foi escolhida por constituir uma zona relevante para a monitorização do fenómeno da desertificação em Portugal. A área de estudo fica localizada na zona interior centro de Portugal continental (Figura 3), abrangendo cerca de 8394.4 km2. Esta área é caracterizada pela presença de um grande corpo de água (a barragem do Alqueva) e de extensas áreas de coberto agrícola (principalmente culturas de sequeiro) no centro. A vegetação esparsa encontra-se essencialmente a Norte; enquanto que a maior concentração de vegetação arbustiva encontra-se a Sul.

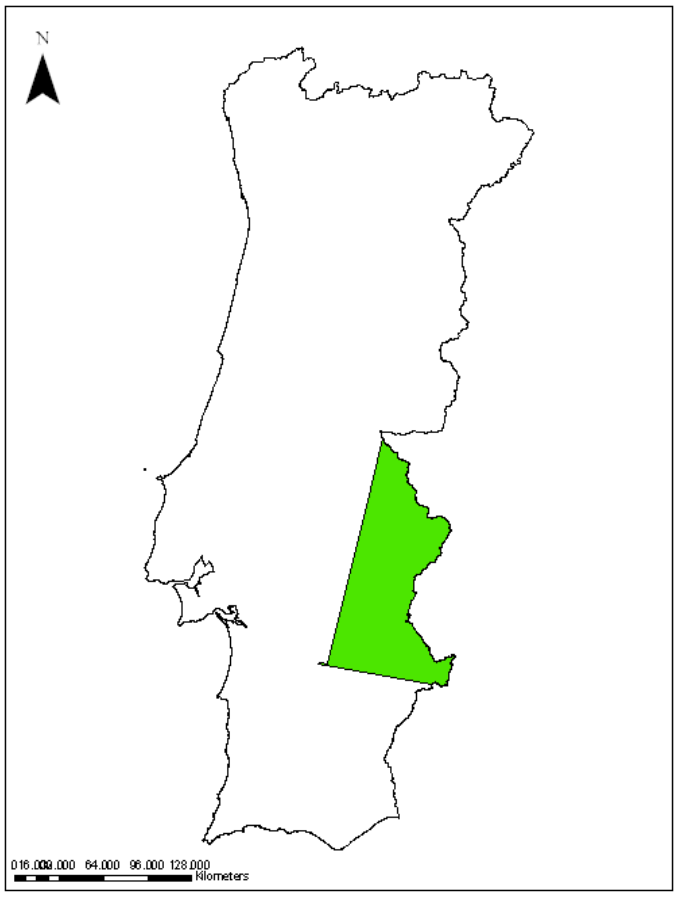

Figura 3 *– Área de estudo a verde.*

3.2 Imagens de Satélite

As imagens utilizadas para o estudo foram adquiridas pelo sensor *Landsat 5* TM (Tabela 1). Estas imagens foram escolhidas pelas seguintes razões: i) por se encontrarem disponíveis gratuitamente (requisito imposto pela ESA), ii) por existir uma cobertura intra-anual para os três países em estudo (Portugal, Brasil e Moçambique) e por iii) já se encontrarem georreferenciadas e ortorrectificadas.

Número da Banda	Comprimento de onda (μm)	Resolução (m)
1	0.45-0.515	30 m
2	0.525-0.605	30 m
3	0.63-0.69	30 m
4	0.75-0.90	30 m
5	1.55-1.75	30 m
6	10.4-12.5	60 m
7	2.09-2.35	30 m
8	0.52-0.9	15 m

Tabela 1 – Características técnicas das bandas da Landsat 5 TM.

Para Portugal foi necessário aplicar uma transformação de coordenadas, uma vez que uma das condições era que o sistema de coordenadas para Portugal fosse o ETRS89 / TM06. Assim, foram seleccionadas duas imagens *Landsat* compreendendo duas datas intra-anuais, Julho (Figura 4) e Novembro de 2009. Estas datas foram escolhidas de modo a representarem dois momentos sazonais diferentes. Desta forma é possível acompanhar parte do comportamento anual do coberto vegetal, nomeadamente as zonas de carácter agrícola. Destas imagens foram utilizadas todas as bandas excepto a sexta banda (banda térmica).

Figura 4 – Imagem Landsat 5 TM da área de estudo (Julho 2009).

Composição colorida: Infravermelho próximo, vermelho, verde.

3.3 Nomenclatura e Subclasses Espectrais

As classes incluídas na nomenclatura são factores importantes para a posterior derivação dos indicadores de desertificação (Panigada et al., 2009). Como dito anteriormente, os mapas LULC podem evidenciar degradação do solo por meio da proporção da ocupação de zonas como áreas urbanizadas, solo nu, áreas ardidas, etc. (Panigada et al., 2009). A nomenclatura adoptada no projecto DWE foi baseada na nomenclatura utilizada no projecto *Land Degradation Assessment in Drylands* (LADA) (Nachtergaele e Petri, 2008), nas recomendações da UNCCD e nos requisitos identificados pelos utilizadores finais dos mapas LULC. A cada uma das classes da nomenclatura utilizada no DWE fez-se corresponder uma ou várias classes definidas no *Land Cover Classification System* (LCCS). Deste modo, é possível relacionar as classes DWE com as classes encontradas no *Globcover* e obter uma caracterização mais completa de cada uma das classes. Na tabela 2 encontram-se definidas as classes LULC utilizadas no projecto DWE, assim como a(s) sua(s) classe(s) correspondente(s) no LCCS.

Classe LULC (DWE)	Código DWE	Código LCCS	Código CLC/COS	Caracterização da classe no LCCS
Urbano	**1**	0010		Áreas contendo no máximo 4% de coberto vegetal durante pelo menos 10 meses por ano. As áreas manipuladas por intervenção

				humana.
Agricultura	**2**	10001 - W8 10025 - W8 10013 - W8		Áreas utilizadas para a agricultura, constituídas por terras aráveis e culturas permanentes .
Floresta	**3**	21445		Áreas ocupadas por conjuntos de árvores florestais resultantes de regeneração natural, sementeira, ou plantação. As árvores devem, no seu conjunto, constituírem um grau de coberto superior ou igual a 30%.
Matos	**4**	21449		Áreas naturais de vegetação espontânea arbustiva, pouco ou muito densa, em que o grau de coberto arbustivo é superior ou igual a 30%.
Pastagem	**5**	21453		Zonas de vegetação herbácea em que esta ocupa uma área

				superior ou igual a 30% da superfície e que se desenvolvem sem adubação, cultivos, sementeiras ou drenagens. Estas áreas podem ser utilizadas de forma extensiva para pastoreio.
Vegetação Esparsa	**6**	20049 20058		Áreas de vegetação esparsa em que a superfície com vegetação arbórea, arbustiva e herbácea ocupa uma área superior ou igual a 10% mas inferior a 30%, estando a restante área sem vegetação.
Áreas Ardidas	**7**	NA		Áreas florestais e/ou naturais e seminaturais afectadas por fogos recentes, ardidas há menos de 3 anos, que na imagem ainda apresentam um aspecto negro. Não inclui áreas que

				demonstrem sinais de regeneração da floresta.
Solo Nu	**8**	0011		Praias, dunas e extensões de areia, seixos ou calhaus rolado em zonas costeiras ou interiores, incluindo o leito de cursos de água com regime torrencial e áreas de solo nu, com coberto vegetal inferior a 10% e sem uso agrícola, florestal ou urbano. Inclui áreas em que a superfície coberta por rocha ocupa uma área superior ou igual a 90%.
Zonas Húmidas	**9**	0005		Zonas húmidas interiores que incluem zonas apaúladas e turfeiras; zonas húmidas litorais que incluem sapais, juncais e caniçais halófitos; salinas e zonas entre-

				marés.
Corpos de Água	**10**	7002 8002		Superfícies de água doce que incluem cursos de água e planos de água, naturais ou artificiais; superfícies de água salgada, que incluem oceanos e/ou água salobra que incluem lagoas costeiras desembocaduras fluviais.

Tabela 2 – Nomenclatura utilizada no projecto DWE. (NA = não se aplica.)

As classes Agricultura e Floresta, tal como estão definidas, são classes com uma grande complexidade espectral, uma vez que incluem diferentes tipos de coberto com respostas espectrais dispares, como por exemplo regadio e sequeiro na classe Agricultura, e floresta de resinosas e floresta de folhosas na classe Floresta. Como consequência, as assinaturas espectrais destas classes não teriam um comportamento unimodal como seria de esperar, mas multimodal, o que iria tornar a aplicação de algoritmos baseados em estimativas de parâmetros estatísticos (e.g. o LDC e o QDC) ineficientes. Assim, a classe Agricultura e a classe Floresta foram decompostas em subclasses espectrais mais puras. A classe Agricultura foi dividida nas classes Regadio e Sequeiro; e a classe Florestas nas classes Floresta de Resinosas, Floresta de Folhosas e Floresta Mista. Estas divisões deram origem à nomenclatura das subclasses espectrais contendo 13 classes (Tabela 3).

Subclasses LULC	Acrónimo das Subclasses	Classes LULC
Urbano	1	Zonas Artificias
Sequeiro	2	Zona Agrícola
Regadio	3	Zona Agrícola
Floresta de Folhosas	4	Zona Florestal
Floresta de Resinosas	5	Zona Florestal
Floresta Mista	6	Zona Florestal
Pastagem	7	Pastagem
Matos	8	Matos
Solo Nu	9	Solo Nu
Zonas Ardidas	10	Zonas Ardidas
Zonas Húmidas	11	Zonas Húmidas
Corpos de Água	12	Corpos de Água
Vegetação Esparsa	13	Vegetação Esparsa

Tabela 3 – Nomenclatura das subclasses espectrais.

3.4 Amostras de Treino e de Teste

Nesta secção apresentam-se os procedimentos adoptados para a recolha

das amostras de treino e de validação dos mapas. Em particular, expõe-se o modo como as amostras de treino foram tratadas de modo a identificar indivíduos anómalos ao treino.

3.4.1 Selecção da Amostra de Treino

Como explicado anteriormente na secção 2, os métodos de classificação supervisionada necessitam de informação *a priori* sobre os indivíduos de cada classe a ser reconhecida. Essa informação é dada na forma de uma amostra de treino. Essa amostra será posteriormente utilizada para estimar os parâmetros estatísticos de cada classe ou com uma listagem de protótipos de cada classe, dependendo do tipo de classificador a ser utilizado. Em qualquer dos casos, a amostra de treino deverá ser suficientemente grande e representativa de cada uma das classes (Mather, 2004). De facto, o número mínimo teórico de indivíduos de treino por classe é de 30d, onde d é o número de atributos (Mather, 2004). Por outro lado, a qualidade do treino está também fortemente dependente da estratégia de selecção da amostra de treino (*pixel* ou polígono) e da autocorrelação espacial, especialmente em imagens de grande resolução (Chen e Stow, 2002).

A estratégia adoptada para a selecção da amostra de treino foi uma recolha determinística por polígonos. Esta estratégia tende a estabelecer o melhor compromisso entre tempo despendido na recolha da amostra e dimensão (Chen e Stow, 2002). Para a recolha de polígonos, realizou-se uma classificação não supervisionada, por meio do algoritmo *k-means,* de modo a agrupamento de *pixels* espectralmente homogéneos. Os polígonos foram, então, desenhados pelo analista dentro dos agrupamentos identificados. Cada polígono foi definido de modo a conter pelo menos 9 *pixels* (janela 3 x 3). Para minimizar a autocorrelacção espacial entre as amostras, os polígonos foram definidos por toda a área de estudo (Figura 5).

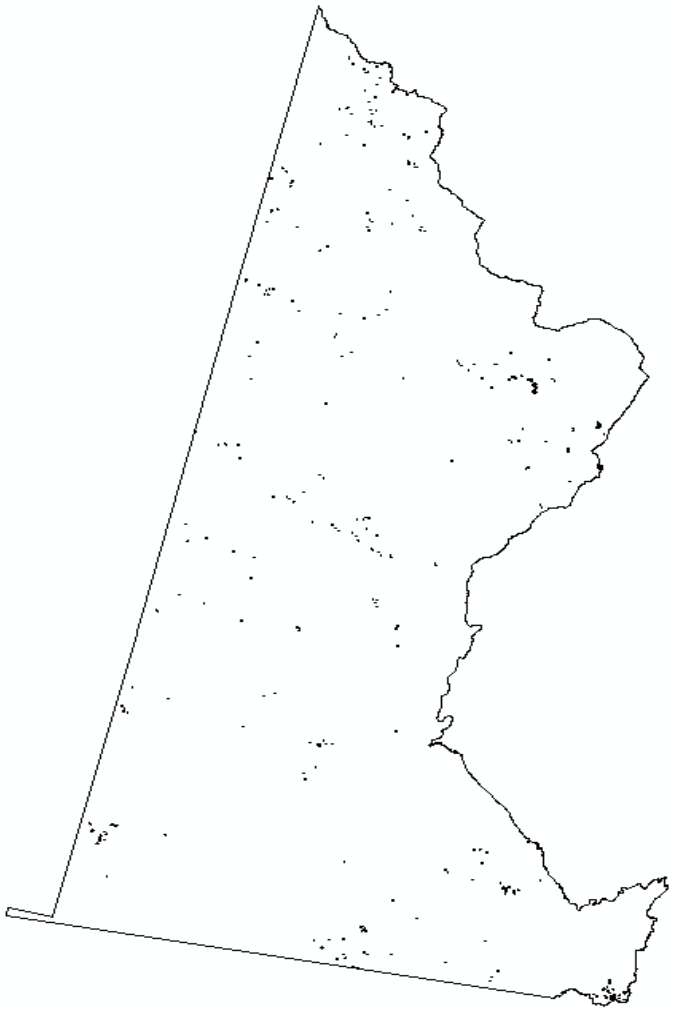

Figura 5 – Distribuição dos polígonos da amostra de treino.

Para auxiliar a interpretação da imagem, o analista teve acesso aos seguintes dados auxiliares, resumidos na Tabela 4.

Nome	Descrição	Data
Divisão Administrativa Oficial de Portugal (CAOP)	Carta com a organização administrativa de Portugal Continental	2009
CORINE Land Cover - CLC00	Mapa de Ocupação do Solo, ano 2000	2002-2005

CORINE Land Cover - CLC06	Mapa de Ocupação do Solo, ano 2006	2007-2009
Mapa de Vinhas	Mapa da distribuição das vinhas em Portugal Continental	1930-2000
Mapa das Zonas Ardidas	Mapa Nacional da Áreas Ardidas dos anos de 1990 a 2008	1990-2008
Fotografias Aéreas	Fotografias aéreas dos anos 1995, 2005 e 2007	1995, 2005, 2007

Tabela 4 – Dados auxiliares utilizados na interpretação dos polígonos para a amostra de treino.

Segundo o CLC06, a área de estudo não contem nem áreas ardidas nem zonas húmidas; esse facto foi confirmado pelo analista que recolheu amostra de treino. Assim, as classes Áreas Ardidas e Zonas Húmidas não irão entrar no mapa final. Para as restantes subclasses espectrais procurou-se recolher 30 polígonos (Tabela 5). Este número foi escolhido por ser o melhor compromisso entre o tempo despendido na recolha da amostra e número total de *pixels* de treino. Contudo, na classe Floresta Mista, não foi possível definir-se esse número de polígonos devido à reduzida ocupação de solo desta classe na área de estudo.

Subclasses Espectrais	Código das Subclasses Espectrais	Classes DWE	Número de Polígonos Recolhidos
Urbano	1	Urbano	30
Sequeiro	2	Agricultura	30

Regadio	3	Agricultura	30
Floresta de Folhosas	4	Floresta	30
Floresta de Resinosas	5	Floresta	30
Floresta Mista	6	Floresta	10
Pastagem	7	Pastagem	30
Matos	8	Matos	30
Solo Nu	9	Solo Nu	30
Áreas Ardidas	10	Áreas Ardidas	0
Zonas Húmidas	11	Zonas Húmidas	0
Água	12	Água	30
Vegetação Esparsa	13	Vegetação Esparsa	30

Tabela 5 – Número de polígonos recolhidos por subclasse espectral.

3.4.2 Identificação de Índividuos Anómalos ao Treino

No presente trabalho, a identificação de indivíduos anómalos ao treino (*outliers*) foi realizada de duas formas: i) identificação estatística, recorrendo às distribuições estatísticas definidas pelas assinaturas espectrais; e ii) identificação por meio de regras periciais, baseadas no conhecimento do comportamento especifico das subclasses espectrais.

3.4.2.1 Identificação Estatística

Apesar do recurso a informação auxiliar, a recolha de amostras de treino para classes LULC livres de indivíduos anómalos é por vezes difícil (Carrão et al., 2010). Os indivíduos das amostras de treino são *pixels* de

um *raster* com diversas imagens e bandas sintéticas, i.e. cada *pixel* é um vector com tantas dimensões quanto o número de bandas consideradas no processo de classificação, $p = [x_1, \ldots, X_d]^T \in \mathbb{R}^d$. A identificação de *outliers* num espaço multidimensional passa pela definição da distância de *Mahalanobis* (Johnson e Wichern, 1998), do seguinte modo: seja C_i a amostra de treino da classe i contendo n indivíduos e seja $x \in C_i$. O quadrado da distância de Mahalanobis entre x e o centro de massa da distribuição da classe i é dada por:

$$d^2(x, u_i) = (x - u_i)^T \Sigma_i (x - u_i) \quad (20)$$

onde u_i é o vector médio eΣ_ia matriz de covariâncias da classe i. O vector médio e a matriz de covariâncias são estimados pelos seus respectivos estimadores de máxima verosimilhança dados pelas equações 21 e 22, respectivamente:

$$\hat{p} = \frac{1}{n} \sum_{j=1}^{n} p_j \quad (21)$$

$$\hat{\Sigma_i} = \frac{1}{n-1} \sum_{j=1}^{n} (p_j - \hat{u_i})(p_j - \hat{u_i})^T \quad (22)$$

Nestas condições, se a classe i tiver um comportamento normal multivariado, então a distância de Mahalanobis é modelada por uma distribuição Qui-Quadrado com d graus de liberdade, $X_d(\alpha)$ (Pestana e Silvio, 2002).

Assim, de modo a identificar indivíduos do treino anómalos à distribuição da classe i, bastará comparar os quadrados das distâncias de *Mahalanobis* amostrais calculadas com a equação 20 com o valor teórico $X_d^2(\alpha)$ tabelado, onde α é o nível de significância do teste. Se

$d^2(x, u_i) > X_d^2(\alpha)$, rejeita-se a hipótese nula que x pertence à classe i, com um nível de confiança de $100(1-\alpha)\%$; caso contrário, a hipótese nula não pode ser rejeitada, pelo que x é mantido na amostra de treino da classe i.

3.4.2.2 Identificação Pericial

A aplicação de regras periciais na amostra de treino é uma forma não-estatística de identificação de indivíduos anómalos à amostra. Estas regras baseiam-se no conhecimento que os peritos possuem sobre os objectos do espaço de classificação e das classes da nomenclatura. No presente caso, o espaço de classificação é uma imagem satélite, os objectos são os seus *pixels* e as classes são as classes LULC.

Para verificar como é que os classificadores se comportavam com esta amostra de treino, foram produzidos diversos mapas preliminares. Foi observado que os classificadores tendiam a cometer muitos erros de comissão nas classes Sequeiro e Pastagem, Urbano e Solo Nu. A amostra de treino foi examinada e concluiu-se que, por exemplo, existiam alguns elementos da amostra de treino da classe Pastagem que poderiam ser classificados como Sequeiro, possivelmente contribuindo para confusão observada. De modo a evitar que um analista tivesse que inspecionar cada um dos indivíduos de treino, foram desenvolvidas regras periciais incidindo sobre as classes Sequeiro, Pastagem, Solo Nu e Urbano recorrendo ao NDVI[6].

As seguintes regras para as classes Sequeiro e Pastagem são:

1. Para todo o x da amostra de Sequeiro, se $NDVI_{Jul}(x) - NDVI_{Nov}(x) > R$, então remover x da amostra de treino;

2. Para todo o x da amostra de Pastagem, se $NDVI_{Jul}(x) - NDVI_{Nov}(x) < S$, então remover x da amostra de treino.

[6]O NDVI foi calculado recorrendo à formula (NIR - RED) / (NIR + RED).

A razão para estas regras está no calendário rural (Ripado, 1991). Segundo esta fonte, no primeiro conjunto de imagens que compõe este *conjunto de dados* (Julho) as zonas agrícolas com sequeiro tendem a possuir um nível de maturação elevado e, portanto, mais vegetação (Figura 6).

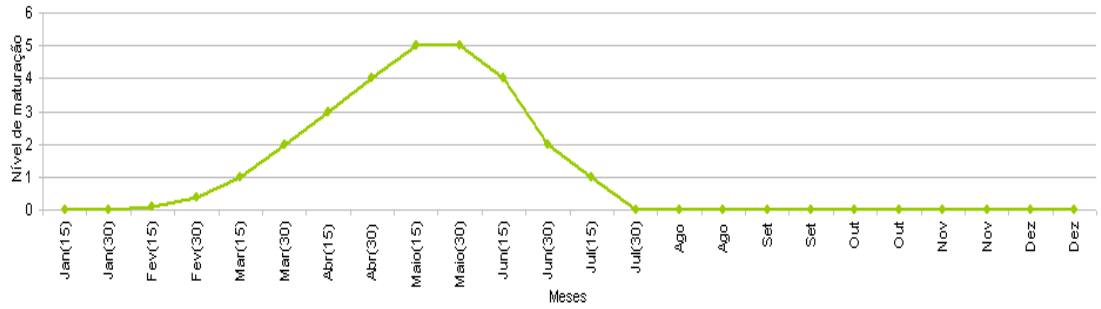

Figura 6 – Comportamento anual do nível de maturação da cultura de sequeiro. Adaptado de Ripado, M.F., Calendário Rural, 1991.

Na segunda data (Novembro), os terrenos são preparados para um novo ciclo de produção, o que passa pela remoção de todo o seu coberto vegetal.

Na classe Pastagem, por outro lado, não é usual a manipulação do coberto vegetal por intervenção humana, pelo que esta classe é caracterizada por ser composta maioritariamente por vegetação herbácea natural. Deste modo, na primeira data (Julho), o esperado é que as zonas de pastagem se encontrem praticamente despidas de qualquer coberto vegetal, devido à falta de irrigação e ao excesso de sol, enquanto que na segunda data se espera que já existam indícios de vegetação natural devido às primeiras chuvas do ano (Novembro).

Nas classes Urbano e Solo Nu, por outro lado, o NDVI tende a não sofrer grandes variações, quer positivas quer negativas, permanecendo razoavelmente constante ao longo do ano. Deste modo, as regras periciais construídas para as classes Urbano e Solo são as seguintes:

3. Para todo o x da amostra de treino da classe Urbano, se

$|NDVI_{Jul}(x) - NDVI_{Nov}(x)| > U$, então remover x da amostra de treino;

4. Para todo o x da amostra de treino da classe Solo Nu, se $|NDVI_{Jul}(x) - NDVI_{Nov}(x)| > T$, então remover x da amostra de treino.

Os valores R e S, das regras 1 e 2, e os valores U e T, das regras 3 e 4, foram determinados experimentalmente por meio da produção de mapas preliminares com diversos classificadores (LDC, QDC e 1-NN) para um excerto da área de estudo. A partir dessas experiências estabeleceu-se que R = 0, S = 0, U = 2 e T = 1.

3.4.3 Selecção da Amostra de Teste e Validação de Mapas

Nesta secção são apresentados os conceitos elementares relativamente à validação de mapas temáticos, assim como o protocolo de validação adoptado para os mapas produzidos.

3.4.3.1 Matriz de erro

A matriz de erro (ou matriz de confusão) compara a classificação realizada por um processo de classificação para um conjunto de objectos do espaço de classificação com os *labels* de referência (Kuncheva, 2004), permitindo observar o modo como os erros de classificação de distribuem ao longo das classes de referência. Estes erros podem ser de dois tipos: erro de comissão, ou erro de omissão (Congalton e Green, 2009). Os erros de comissão são erros de classificação caracterizados pela atribuição de uma classe errada a um determinado objecto; o erro de omissão, por outro lado, consiste na não-inclusão de um determinado objecto na sua verdadeira classe (Congalton e Green, 2009). A matriz de erro pode, então, ser disposta do seguinte modo:

		Referência						
		C1	C2	…	Cj	…	Cn	
Classificação	C1	N11	N12	…	N1j	…	N1n	U1
	C2	N21	N22	…	N2j	…	N2n	U2
	…	…	…	…	…	…	…	…
	Ci	Ni1	Ni2	…	Nij	…	Nin	Ui
	…	…	…	…	…	…	…	…
	Cn	Nn1	Nn2	…	Nnj	…	Nnn	Un
		P1	P2	…	Pj	…	Pn	E.G.

Tabela 6 – Matriz de erro.

onde Nij representa o número de objectos classificados como Ci mas que na referência são Cj. A linha a cor-de-laranja representa a **exactidão do produtor**, definida por:

$$P_i = \frac{N_{ii}}{N_{.i}} \quad (23)$$

onde $N_{.i}$ é a soma dos valores na coluna i. A exactidão do produtor Pi indica a proporção de índividuos da classe de referência Ci que foram correctamente classificados. A exactidão do produtor é, então, uma medida para o erro de omissão (Congalton e Green, 2009). A coluna i mostra o modo como a classe de referência Ci se encontra distribuída pela classificação.

A coluna a cor-de-laranja representa da **exactidão do utilizador**, definida por:

$$U_i = \frac{N_{ii}}{N_{i.}} \quad (24)$$

onde $N_{i.}$ é a soma dos valores na linha i. A exactidão do utilizador Ui indica a proporção de índividuos classificados como membros da classe

Ci que foram correctamente classificados, e deste modo a exactidão do utilizador é uma medida do erro de comissão (Congalton e Green, 2009). Assim, a linha i mostra a distribuição dos objectos classificados como Ci sobre as classes da referência.

As iniciais E.G. significam **exactidão global**, que é definida por:

$$EG = \frac{N_{ii}}{N} \quad (25)$$

onde N é o número total de indivíduos na amostra de referência. A exactidão global indica a proporção de indivíduos da amostra de referência que foram correctamente classificados e, portanto, é uma medida da qualidade global da classificação (Congalton e Green, 2009).

3.4.3.2 Amostra de Teste

A dimensão da amostra de teste é o factor com maior impacto na avaliação da qualidade do mapa (Dicks e Lo, 1990). A dimensão da amostra de teste deve ser tal que, assim como a exactidão global, com uma determinada incerteza associada. Para mais, a dimensão da amostra de teste deve ser suficientemente grande para permitir a realização de um teste de hipóteses sobre a exactidão temática do mapa.

A regras para determinar o número mínimo de elementos de uma a amostra de teste têm sido baseadas no modelo probabilístico binomial (Ginevan, 1978; Aronoff, 1982). A partir deste modelo é possível construir tabelas que relacionam o número mínimo de elementos de teste com o risco do produtor e com o risco do utilizador[7]. A tabela A6 em Aronoff (1985) mostra que para um risco de produtor e um risco de

[7]O risco do produtor é a probabilidade do teste de hipóteses realizado rejeitar um mapa com uma exactidão temática global superior a um determinado valor fixo à partida, e.g. 90%. O risco do utilizador, por outro lado, é a probabilidade do teste de hipóteses aceitar um mapa temático com uma exactidão temática global inferior a um determinado valor fixo à partida, usualmente 85% (Ginevan, 1978).

utilizador de 10%, para valores de 90% e 85%, respectivamente, a dimensão mínima da amostra de treino é de 288 (~ 300) elementos. Este número indica-nos a dimensão mínima da amostra. Contudo, é necessário distribuir esse valor por cada uma das classes, de modo a que a exactidão do produtor e a exactidão do utilizador sejam estimadas com igual incerteza máxima. Prova-se que (Cochran, 1977) a incerteza da estimativa da exactidão do produtor e da exactidão do utilizador, d, é dada por:

$$d = z_{1-\alpha/2}\sqrt{\frac{1}{4n}} \quad (26)$$

onde $z_{1-\alpha/2}$ é o quantil da distribuição normal para os $100(1-\alpha)\%$. Para um nível de confiança de 95%, se o valor da incerteza for fixo nos 0.1, então n é aproximadamente igual a 96 elementos. Portanto, para que a exactidão do produtor e a exactidão do utilizador sejam estimadas com uma incerteza máxima de 0.1, são necessário 96 elementos de teste por classe, o que totaliza uma amostra de teste com 960 indivíduos. Uma amostra dessa dimensão requer muito tempo de análise, pelo que se encontrou um compromisso entre a incerteza da estimativa e o tempo de recolha de 50 elementos por classe, o que implica uma incerteza de aproximadamente 0.14.

Para o lançamento dos elementos da amostra de teste, o CLC06 foi reclassificado nas classes da nomenclatura DWE. A unidade amostral utilizada foi o *pixel*, pelo que a amostra de teste é composta por 500 *pixels* (Figura 7).

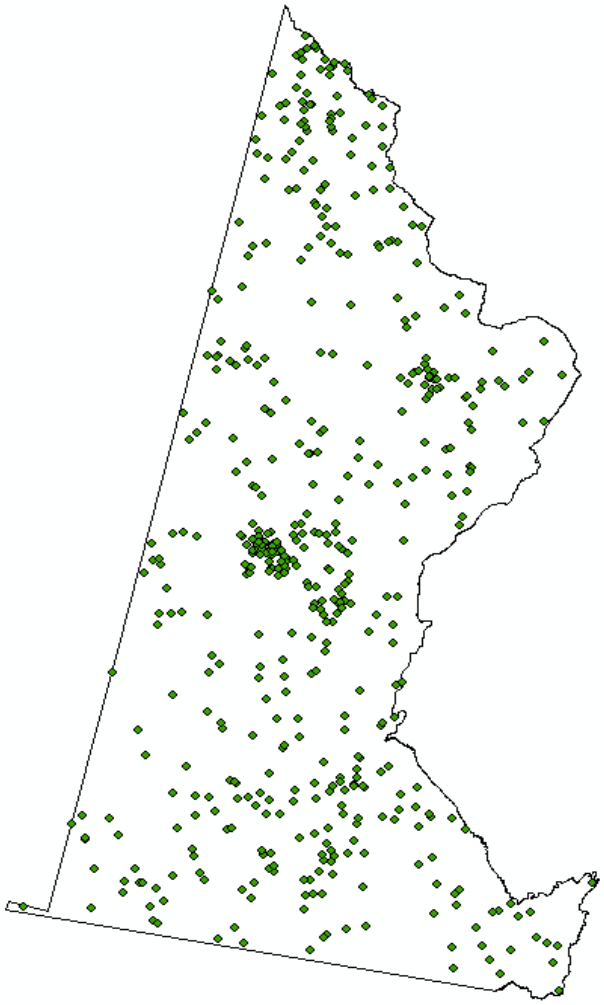

Figura 7 – Distribuição dos elementos da amostra de teste pela área de estudo.

Para a interpretação dos elementos da amostra de teste, foram utilizados os mesmos dados auxiliares que na recolha da amostra de treino (ver secção 3.4.1).

3.4.3.3 Regra de Concordância

No processo de validação de mapas temáticos, existem erros resultantes de factores externos à classificação. De um modo geral, estes erros encaixam-se em duas categorias: erros de coregisto e erros de interpretação (Foody, 2002). Os erros de coregisto podem ter inúmeras fontes, como por exemplo, erros no processo de georreferenciação. O

segundo tipo de erro provém do facto da interpretação de imagens não ser um processo totalmente objectivo, sendo por vezes impossível atribuir uma só classificação ao ponto ou ao polígono utilizados no processo de validação do mapa. Uma regra de concordância que define o acordo que não considere estes dois tipos de erro, irá inclui-los na contabilização dos erros, resultando numa estimativa irrealista da exactidão temática do mapa (Foody, 2002). Por exemplo, se para cada elemento da amostra de teste for atribuída uma só classificação, nem a ambiguidade na interpretação nem a ambiguidade posicional serão considerados (Foody, 2002). Stehman e Czaplewsky (1998), Foody (2002), Wulder et. al (2006), entre outros, defendem que a validação de mapas temáticos resultantes da classificação de imagens de satélite, como Landsat e SPOT, deve ser realizada recorrendo a uma região de suporte (e. g. uma janela de 3 x 3 *pixels*) centrada nos pontos de referência, e que cada ponto contenha dois *labels* alternativos. Deste modo, a região de suporte procura mitigar a ambiguidade posicional e os *labels* alternativos a ambiguidade da interpretação do analista.Nestas condições, um acordo entre o mapa e a referência ocorre sempre que na região de suporte exista pelo menos um *label* igual ao primeiro *label* alternativo ou igual ao segundo; caso contrário, o erro é contabilizado no primeiro *label*. A dimensão da região de suporte usualmente aplicada tem sido a janela 3 x 3 (Zhu et al, 2000; Stehman e Czaplewsky, 2003; Wikham et al, 2004), por ser suficientemente larga para mitigar a ambiguidade posicional, mas não excessivamente larga para incluir uma grande extensão de terreno, tornando o processo de validação pouco significativo. Assim, para avaliar a variabilidade no interior das regiões de suporte, procedeu-se à classificação dos nove *pixels* das janelas 3 x 3 centradas nos pontos de referência e determinou-se o número classes distintas no seu interior. Conclui-se que cerca de 75% das regiões continha no máximo três classes diferentes; tendo as restantes quatro ou cinco, e somente 1% (5 regiões de suporte) continham seis classes (Figura 8). Conclui-se assim que a janela 3 x 3 constitui uma região de suporte válida para o presente caso.

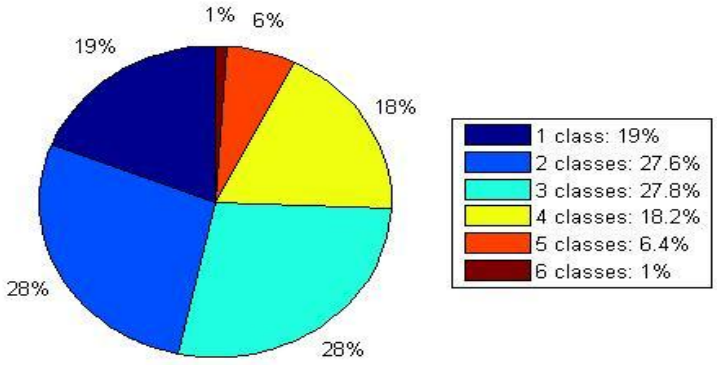

Figura 8 – Análise à variabilidade interna nas regiões de suporte. (Classificador utilizado 1-NN. Os resultados são análogos para as classificações realizadas pelos outros classificadores.)

Deste modo, a validação do mapa é realizada com uma amostra de 500 *pixels* (50 por classe da nomenclatura DWE), cada um com dois *labels* associados e onde a regra de concordância construída por meio da definição de uma região de suporte de 3 x 3 *pixels,* centrada em cada *pixel* da amostra de teste. Uma concordância entre o mapa e a amostra de teste existe sempre que numa região de suporte exista um *pixel* com um *label* igual a um dos *labels* do respectivo *pixel* da amostra de teste.

3.5 Avaliação da Qualidade dos Classificadores Simples

A avaliação da qualidade dos classificadores não é uma tarefa totalmente objectiva, existindo diversos modos de se proceder a essa avaliação. De facto, Cihlar et al. (1998) propõe seis variáveis: exactidão, reproducividade, robustez, capacidade em utilizar toda a informação disponível, aplicabilidade e objectividade. Contudo, na realidade, não existe nenhum classificador que satisfaça todas as seis condições propostas, devido às condições específicas do problema em que são aplicados (Lu e Weng, 2004). Assim, de modo a simplificar o processo

de avaliação de classificadores, propõem quatro factores: exactidão da classificação, recursos computacionais, estabilidade do algoritmo e robustez ao ruído nos dados. No presente estudo, os classificadores foram comparados tendo em conta i) o volume de treino que necessário para obterem a sua exactidão máxima, ii) a robustez ao ruído, iii) a exactidão da classificação e iv) os requisitos computacionais, nomeadamente o intervalo de tempo necessário para realizar a classificação. Para avaliar o primeiro factor, procedeu-se à construção das curvas de aprendizagem; a resistência ao ruído é avaliado pelas curvas de robustez e a exactidão da classificação é avaliada pelo protocolo de validação do mapa descrito na secção 3.4.3.3.

3.5.1 Cross-Validation

O método mais usual para a estimação da precisão global de um algoritmo de classificação é o *cross-validation* (Hastie, et al. 2009). A situação ideal seria recolher três conjuntos de dados: o treino, a validação e o teste. O classificador seria treinado com o treino, afinado com a validação e avaliado com o teste. Porém, a recolha rigorosa de três conjuntos de dados nem sempre é possível (por que os dados são escassos) ou impraticável (por que é uma tarefa que exige muito tempo).

Uma solução para se ultrapassar esta limitação, consiste em treinar o classificador com a amostra de treino e recorrer à mesma amostra para estimar a exactidão global, este método é conhecido por substituição (Kuncheva, 2004). Contudo, desta abordagem resultaria uma estimativa excessivamente optimista (Webb, 2002). De facto, certos classificadores (e.g. *k-nearest neighboor* e árvores de classificação) têm tendência para "memorizarem" o treino (Kuncheva, 2004), o que implica que se o treino fosse utilizado também para teste, estes classificadores tenderiam a obter resultados com exactidões globais muito elevadas, mas pouco realistas.

O método do cross-validation procura ultrapassar esse problema por meio da partição da amostra de treino (Hastie et al., 2009). Isto é, o método *cross-validation* (conhecido também por K-*fold cross-validation*) consiste na partição aleatória do treino em K partes aproximadamente iguais e na aplicação de um ciclo de treino e teste para cada uma das partes, ou seja: a primeira parte é fixa para teste do classificador e as restantes K-1 utilizadas no seu treino, daí é determinado o primeiro valor da estimativa da exactidão global, por meio da proporção de indivíduos do teste correctamente classificados. Depois, a segunda parte é fixa para teste e as restantes K-1 são utilizadas para treino, donde é calculado o segundo valor da estimativa da exactidão global, determinado da mesma forma que anteriormente. O procedimento é repetido desta forma até à K-ésima parte. No final, é determinada a estimativa final da exactidão global, calculada pela média dos valores da estimativa da exactidão global obtidos em cada parte. Deste modo, o método cross-validation procura utilizar a variabilidade interna existente na amostra de treino para obter uma estimativa mais realista da exactidão global do classificador (Hastie et al., 2009).

A questão que se coloca é, que valor atribuir a K? De um modo geral, é aconselhado escolher o valor de 5 ou de 10 por ser um bom compromisso entre a qualidade da estimativa e as exigências computacionais (Fukunaga, 1991; Webb, 2002; Hastie et al., 2009). De facto, se K = 2 (método conhecido por *hold-out*), a estimativa tenderá a ser bastante pessimista (Kuncheva, 2004). Por outro lado, se K for igual ao número de indivíduos do treino (método conhecido por *leave-one-out*), o resultado tende a ser bastante realista, no entanto, os requisitos computacionais são consideráveis.

3.5.2 Curvas de Aprendizagem

As curvas de aprendizagem relacionam o volume de treino (i.e. o número de indivíduos de treino por classe) com a exactidão global da classificação (Hastie et al., 2009). Para a construção das curvas de

aprendizagem, o seguinte procedimento foi realizado 10 vezes: primeiro, são recolhidas 16 amostras de treino, a primeira com 5 elementos, a segunda com 10, a terceira com 15, …, e a última amostra com 80. Os indivíduos a comporem estas amostras de treino são recolhidos aleatoriamente da amostra de treino inicial. Depois, para cada uma das amostras de treino, é realizado um teste *5-fold cross-validation* para cada classificador a ser comparado e é determinada a exactidão global do teste. No final das 10 repetições, é determinada a exactidão global média de cada amostra de treino.

3.5.3 Curvas de Robustez

As curvas de robustez mostram a relação entre a quantidade de ruído e a exactidão global. O procedimento adoptado para inserir ruído na amostra de treino consistiu em repetir o seguinte processamento 5 vezes: primeiro, a amostra de treino de cada classe é dividida em duas partes, uma contendo 80% dos elementos e outra contendo os restantes 20%. A primeira será utilizada como treino, chamemos-lhe treino 0, e irá sofrer degradações sucessivas de ruído; a segunda será utilizada como teste. Segundo o seguinte procedimento é repetido um determinado número de vezes: o treino 0 é degradado com a introdução de ruído nos seus dados dando origem ao treino 1. Os classificadores são, então, treinados com o treino 1 e testados com o teste, donde se determina a exactidão global. Seguidamente, a amostra treino 1 é degradada com mais ruído dando origem à amostra treino 2; aí os classificadores são treinados com o treino 2 e testados com o teste, determinando a exactidão global na segunda iteração. No final, é determinada a exactidão global média em cada interacção.

3.5.4 Comparação Experimental de Classificadores

Quando se fala em comparação experimental de classificadores usualmente associa-se à sequência de testes e no conjunto de conjunto

de dados utilizados para treinar, testar, optimizar e seleccionar classificadores (Kunchuva, 2004). Contudo, estas experiências devem ser conduzidas com algumas linhas de orientação em mente, porque, mostra a experiência, os resultados podem ser facilmente enviesados, mesmo sem intenção do analista. De facto, segundo Kuncheva (2004), para qualquer caso de aplicação, existe um conjunto de dados no qual o classificador candidato é superior a qualquer um dos seus classificadores concorrentes.

O teste deve ser composto por objectos que não participaram no processo de treino. Só assim é possível avaliar a capacidade de generalização de um classificador. (A capacidade de generalização de um classificador é a sua capacidade de classificar correctamente objectos que não se encontram no treino.) Esta situação levanta uma dificuldade acrescida na comparação de classificadores. Esta dificuldade é conhecida por ***overtraining*** (também conhecido por *overfitting*); um classificador dir-se-á que está em *overtraining* quando a sua regra de classificação se encontra extremamente adaptada às características particulares do treino, o que torna a generalização mais difícil. Neste modo, o *overtraining* implica a perda de generalização. Isto é particularmente importante quando um classificador possui parâmetros de entrada que necessitam de ser optimizados (e.g. o número de vizinhos no classificador k-NN). Usualmente, os parâmetros dos classificadores estão fortemente relacionados com a estrutura dos dados e a sua modelação é difícil (Hastie et al, 2009). O processo mais praticado para determinar a parametrização óptima, passa pela realização de uma sequência de testes em *cross-validation* para cada parametrização possível, seleccionando posteriormente aquela que produzir a precisão global máxima (Hastie et al, 2009).

O problema desta abordagem é que força o classificador a ajustar-se demasiadamente bem à estrutura dos dados de treino (Fukunaga, 1991; Webb, 2002). Como consequência, o classificador tende a obter resultados excelentes nos testes com dados conhecidos, mas tornar-se-á

ineficiente quando confrontado com dados do espaço de classificação que são suficientemente diferentes daqueles que se encontram no treino. Este fenómeno encontra-se usualmente associado a classificadores sofisticados, i.e. que permitem a definição de fronteiras de decisão com geometrias complexas, como é o caso das Redes Neuronais, dos classificadores não-paramétricos e das Árvores de Classificação. Classificadores mais simples, como o LDC, usualmente não sofrem deste problema; contudo, estes classificadores perdem a sua capacidade de generalização quando o espaço de classificação apresenta uma disposição onde existe uma elevada sobreposição de indivíduos de diferentes classes. Na Figura 9 apresenta-se um exemplo comparativo entre a definição de fronteiras de decisão com geometria complexa, definida pelo classificador 3-NN, e definição de fronteiras de decisão simples, definida pelo LDC.

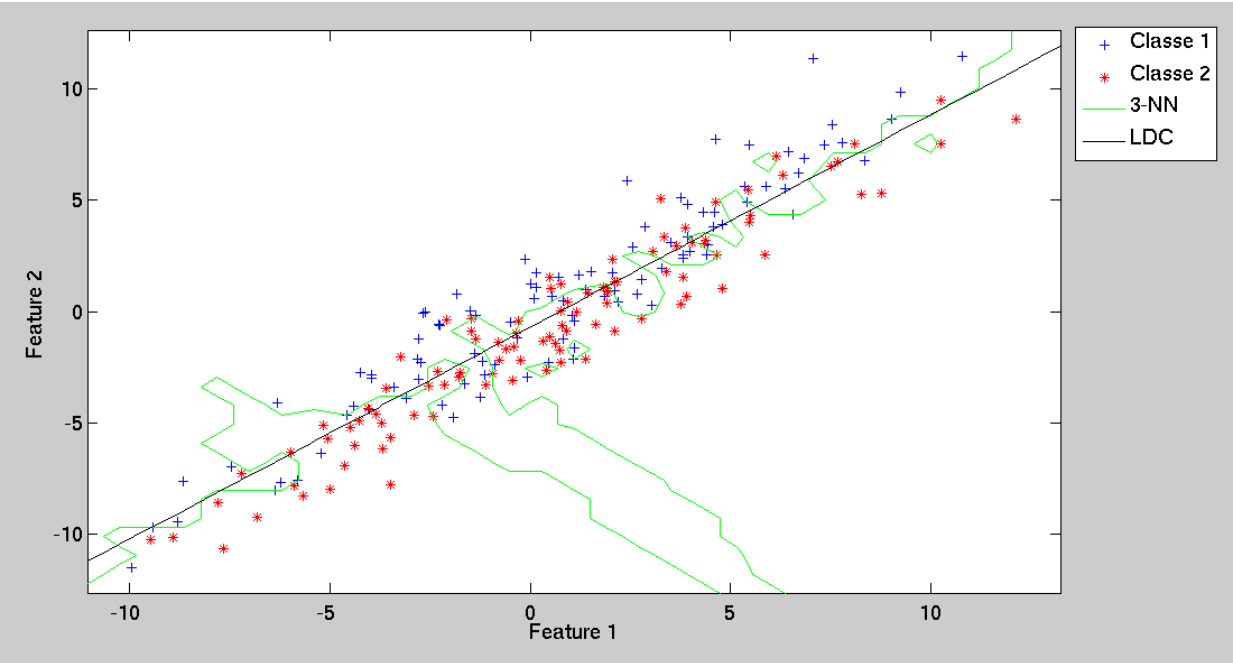

Figura 9 – Comparação da definição de fronteiras: LDC vs. 3-NN.

Em última análise, a performance de um classificador será função do analista (Kuncheva, 2009; Hastie et al, 2009), quer seja pela definição das amostras de treino e de teste, quer pela selecção da parametrização

a ser utilizada, quer pela sequência de testes a serem aplicados aos classificadores. Procurou-se seguir as seguintes linhas de orientação para a comparação dos classificadores sugeridos por Kuncheva (2004) de modo a tornar a comparação dos classificadores mais objectiva: a) seleccione o procedimento e o conjunto de treino e mantenha-o fixo durante todo o treino; b) treine os classificadores candidatos nas mesmas condições; c) garanta que o teste nunca é "observado" por nenhum outro classificador durante o treino; d) se um classificador necessita de parametrização, compare primeiro o classificador com as diferentes parametrizações e só depois com os classificadores concorrentes.

3.5.4.1 Teste de McNemar

O teste de McNemar não é tanto um teste para avaliar a qualidade de um classificador, mas antes um teste para comparar as discordâncias entre dois classificadores. Suponhamos dois classificadores, C1 e C2, com exactidão global de 85% e 90%, respectivamente. Será que esta diferença é significativa? Será que os dados são suficientes para afirmar que o classificador C2 é melhor que o classificador C1? Para responder a esta pergunta, existem diversos métodos estatísticos, entre eles o teste de McNemar, o teste da diferença entre proporções, o teste de Cochran, entre outros (Kuncheva, 2004; Hastie et al., 2009). Contudo, no presente texto, apenas o primeiro teste será abordado, por ser simples e usualmente fiável (Kuncheva, 2004).

Seja N_{11} o número de indivíduos da amostra de teste que foram correctamente classificados pelos dois classificadores, C1 e C2; N_{10} o número de indivíduos correctamente classificados por C1 e erroneamente classificados por C2; N_{01} o número de índividuos erroneamente classificados por C1 mas correctamente classificados por C2; e N_{00} o número de indivíduos mal classificados pelos dois classificadores (Tabela 6). Nestas condições, $N_{11} + N_{10} + N_{01} + N_{00} = N$, onde N é o número total de índividuos do teste.

	C2 correcto	C2 errado
C1 correcto	N_{11}	N_{10}
C1 errado	N_{01}	N_{00}

Tabela 7 – Matriz de contagem para o teste de McNemar

O teste de McNemar responde à pergunta, será que as discordâncias entre os dois classificadores são estatisticamente suficientes para concluir que os classificadores são diferentes? Nesta medida, apenas o número de discordâncias será utilizado no teste de McNemar. Prova-se (Kuncheva, 2004) que a quantidadex^2, definida por:

$$x^2 = \frac{(|N_{01}-N_{10}|-1)^2}{N_{01}+N_{10}} \quad (24)$$

é modelada por uma distribuição Qui-Quadrado com um grau de liberdade, admitindo a hipótese nula de que os classificadores são estatisticamente iguais. Nestas condições, se x^2 for superior ao valor tabelado χ^2 para um nível de significância de α, poder-se-á afirmar que as diferenças encontradas nos classificadores são significativas, com uma incerteza deα, pelo que se pode afirmar que o classificador com maior precisão global superou o classificador concorrente. Caso contrário, as discordâncias não serão suficientes para dar preferência a um classificador em detrimento do outro, uma vez que essas diferenças podem ser explicadas em condições de aleatoriedade.

3.6 Definição do Classificador Composto

Nesta secção é apresentada a arquitectura do classificador proposto

para o presente trabalho. Na primeira subsecção são apresentados os conceitos da teoria da informação que irão ser utilizados posteriormente no classificador e noutras fases da metodologia. Na segunda subsecção apresenta-se a transformação de Fisher que será uma componente fundamental na implementação do algoritmo de classificação composto. Na última subsecção é, então, apresentada a estrutura do classificador proposto.

3.6.1 Medidas Informativas de Proximidade Espectral

Com medidas informativas de proximidade espectral procura-se avaliar a semelhança entre *pixels* ou assinaturas espectrais no espaço multidimensional. A diferença entre as medidas de proximidade, diga-se não-informativas, como é o caso da distância euclidiana e a distância de Manhattan, e as medidas informativas, está na introdução do conceito de entropia informativa.

A **entropia informativa**, ou entropia de Shannon, é uma medida de imprevisibilidade associada a uma variável aleatória (Cover e Thomas, 2006). A entropia informativa pode ser definida do seguinte modo (Shannon, 1948): se X é uma variável aleatória contínua, então a entropia informativa de X, H(X), é dada por,

$$H(X) = -\int_{-\infty}^{+\infty} p(x)\log p(x)dx \quad (25)$$

onde p é a função densidade de probabilidade (fdp) associada a X. O conceito de entropia informativa pode ser alterada de modo a passar a ser uma medida associada um par de variáveis aleatória, por meio da **divergência de Kullback-Leibler**.

Suponhamos que X e Y são duas variáveis aleatórias contínuas com fdp p e q, respectivamente. A divergência de Kullback-Leibler (ou ainda, distância de Kullback-Leibler ou entropia relativa) é definida do seguinte

modo (Cover e Thomas, 2006):

$$D_{KL}(X,Y) = \int_{-\infty}^{+\infty} p(x)\log\frac{p(x)}{q(x)}dx \quad (26)$$

A divergência de Kullback-Leibler pode ser entendida como a quantidade de informação necessária para se obter a informação de Y usando a informação de X (Kullback, 1978); e nesta forma, pode ser entendida como uma "distância" ou medida de proximidade entre duas variáveis aleatórias, uma vez que a divergência de Kullbck-Leibler só é nula quando as variáveis aleatórias forem iguais, e quanto menor for o seu valor, mais "próxima" de Y se encontra X (Cover e Thomas, 2006). Contudo, a divergência de Kullback-Leibler não é uma verdadeira distância, uma vez que não satisfaz a condição de simetria (i.e. de um modo geral, $D_{KL}(X,Y) \neq D_{KL}(Y,X)$, nem a desigualdade triangular (Kullback, 1978). Contudo, a divergência de Kullback-Leibler pode ser manipulada de modo a originar uma métrica informativa entre variáveis aleatórias. Nas mesmas condições da definição da divergência de Kullback-Leibler, define-se a **distância de Jensen-Shanon** entre X e Y pela seguinte equação:

$$D_{JS}(p,q) = \frac{1}{2}D_{KL}(X,M) + \frac{1}{2}D_{KL}(Y,M) \quad (27)$$

onde M é a média das variáveis X e Y, i.e. $M = \frac{X+Y}{2}$.

Pergunta-se, então, de que modo os conceitos anteriormente apresentados podem melhorar o processo de classificação?

A distância (em termos estatísticos) entre duas populações é entendida como a dificuldade ou facilidade de discriminá-las (Kullback e Leibler, 1951). A ideia de utilizar distâncias de carácter estatístico para avaliar a separabilidade de classes tem sido recorrente em Detecção Remota,

como é o caso da distância de Battacharya (Lu e Weng, 2004). Em particular, o uso da medida de entropia informativa foi utilizada por Chang (2000) para avaliar a entropia de um *pixel* hiperespectral e, posteriormente, criar duas medidas de semelhança entre *pixels* baseada na divergência de Kullback-Leibler para avaliar a semelhança entre assinaturas espectrais. Mas até então as medidas de separabilidade entre classes têm sido utilizadas como um instrumento exterior ao algoritmo de classificação.

No presente estudo, a divergência de Kullback-Leibler será utilizada para identificar as classes mais próximas de uma dada classe e, portanto, mais prováveis de criar confusão no classificador. A razão para a utilização da divergência de Kullback-Leibler e não, por exemplo, a distância de Jensen-Shannon, advém do facto da divergência de Kullback-Leiber ser assimétrica. A experiência mostra que, na classificação de imagens de satélite, existem classes que são confundidas com outras, mas não o contrário, i.e. a frequência com que uma classe A é confundida com a classe B, não é usualmente igual à frequência com que a classe B é confundida com a classe A. Esta observação mostra que o fenómeno de confusão entre classes LULC durante a classificação tende a ser assimétrico. Deste modo, pretende-se identificar essas assimetrias com a divergência de Kullback-Leibler. Para o cálculo da divergência de Kullback-Leibler entre duas classes, será utilizado o seguinte resultado teórico (Zhou e Chellappa, 2004): sejam X e Y duas variáveis aleatórias contínuas multivariadas e normais, parametrizadas do seguinte modo: $X \sim N(\mu_0, \Sigma_0)$e $Y \sim N(\mu_1, \Sigma_1)$. Então, a divergência de Kullback-Leiber de X para Y é dada por,

$$D_{KL}(X,Y) = \frac{1}{2}\left\{\log\frac{|\Sigma_1|}{|\Sigma_0|} + tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_0^{-1}(\mu_1 - \mu_0) - N\right\} \quad (28)$$

onde | . | é a função determinante, tr (.) é a função traço e N é o número de dimensões das variáveis. Os vectores médio e as matrizes de

covariância são estimadas recorrendo aos respectivos estimadores de máxima verosimilhança.

A distância de Jensen-Shannon, por outro lado, será utilizada como medida de discriminação global de um conjunto de classes espectrais, com a finalidade de determinar as condições em que as classes se encontram mais afastadas umas das outras. Em particular, a determinação do melhor nível de significância para a aplicação da identificação estatística de indivíduos anómalos à amostra de treino será escolhido em função da distância de Jensen-Shannon média entre classes.

3.6.2 Análise Discriminante de Fisher

A Análise Discriminante de Fisher (FDA – *Fisher Discriminant Analysis*), tal como a análise em componentes principais e a análise factorial, encontra-se na família das transformações lineares de carácter estatístico que permitem a redução da dimensionalidade do problema. O que distingue a FDA das restantes transformações desta família é a optimização em termos de separabilidade entre classes, i.e. enquanto que as restantes transformações não tomam em consideração as classes definidas nos dados, na FDA a identificação das classes é fundamental. A FDA tem sido utilizada extensivamente em problemas de reconhecimento de padrões, nomeadamente no reconhecimento de rostos (Webb, 2002; Bow, 2002), mas também na classificação de imagens de satélite (Park et al., 2007; Akgun et al., 2009).

A FDA baseia-se em dois tipos de variância: na **variância interna** ou intra-classe, e a **variância externa** ou entre-classes. A variância interna é uma medida de variabilidade que indica a coesão média das classes do problema; é definida pela seguinte equação:

$$\Sigma_w = \sum_{i=1}^{n} P_i \Sigma_i \quad (29)$$

onde P_i é a probabilidade *a priori* associada à classe i e Σ_i é a matriz de covariância dessa classe. A variância externa, por outro lado, indica a variabilidade dos centros de massa das distribuições de cada classe; é dada pela seguinte equação:

$$\Sigma_b = \sum_{i=1}^{n} P_i (M_i - M_0)^T (M_i - M_0) \quad (30)$$

onde M_i é o vector valor médio da classe i eM_0o vector médio de todos os$M_i$ponderado pela probabilidade *a priori* de cada classe.

A FDA consiste, então, em determinar uma matriz w (chamada **transformação de Fisher**) tal que $Y = w^T X$ em que X é o espaço de classificação inicial com, digamos, D dimensões, e Y é o espaço de classificação transformado com, no máximo, C-1 dimensões, onde C é o número de classes do problema. Adicionalmente, impõe-se que a transformação seja tal que w maximize a separabilidade entre as classes, definida pela razão entre a variância externa e a variância interna do seguinte modo:

$$\frac{w^t \Sigma_b w}{w^t \Sigma_w w} \quad (31)$$

O procedimento para a determinação desta transformação linear passa por decompor em, vectores próprios, a matriz $\Sigma_w^{-1}\Sigma_b$ e, seguidamente, seleccionar os d vectores próprios associados aos d maiores valores próprios, onde $d \leq C - 1$ é o número de dimensões desejado para o espaço transformado Y. A matriz w é, então, a matriz de dimensão D x d formada pelos d vectores próprios anteriormente seleccionados, dispostos em colunas. Estes d vectores são chamados de **componentes principais**.

3.6.3 Arquitectura do classificador composto

Neste trabalho, o classificador proposto é um classificador composto, do tipo cooperativo em sequência. Este classificador é composto por dois classificadores singulares (e.g. LDC, QDC, K-NN, etc.) dispostos em sequência e uma regra de selecção de classes. O primeiro classificador será denominado **classificador genérico** e o segundo classificador **classificador específico**. A regra de selecção de classes consiste na selecção das classes mais prováveis de confusão com uma dada classe. Intuitivamente, as classes mais prováveis de confusão são aquelas que se encontram mais próximas. Esta regra será baseada na divergência de Kullback-Leibler. Posteriormente, como as classes mais prováveis de confusão tendem a estar mais próximas umas das outras, é determinada uma transformação de Fisher por meio dos dados de treino das classes seleccionadas pela regra de selecção. O objecto a ser classificado é projectado para o espaço de máxima separabilidade definido pela transformação de Fisher calculada e a sua classificação é realizada nesse espaço.

O algoritmo de classificação consiste na seguinte sequência de passos (Quadro 4):

Algoritmo 4 – Arquitectura genérica do classificador proposto

Legenda:

i) Ci, conjunto das classes mais próximas da classe i;

ii) Ti, transformação de Fisher associada à classe i;

iii) p*, elemento do espaço de classificação p transformado pela transformação de Fisher.

Input:

i) Espaço de classificação, E;

ii) Classificador genérico, A;

iii) Classificador especifico, B;

iv) Regra de selecção de classes, R;

Output:

Espaço de classificação E classificado.

Procedimento:

1. Para cada classe i,

1. 1. Determinar as classes mais próximas da classe i, Ci, de acordo com R;

1. 2. Determinar a transformação de Fisher, Ti, com as amostras de treino de Ci;

2. Para cada $p \in E$:

2. 1. Classificar p com A (*label* resultante c);

2. 2. Transformar p com a transformação Tc (resultado da transformação, p*);

2. 3. Reclassificar p* com B usando apenas as classes de Cc.

Quadro 4 – Algoritmo do classificador composto.

4 Resultados e Discussão

Nesta secção apresentam-se os resultados obtidos com a metodologia utilizada. Apresenta-se primeiro a aplicação da distância de Jensen-Shannon para a definição do melhor *conjunto de dados* para a discriminação da classes. A mesma medida é aplicada depois para identificar o melhor nível de confiança para a separabilidade das classes. Seguidamente são apresentados os resultados da comparação dos classificadores e da aplicação do classificador composto.

4.1 Conjunto de dados Utilizado

Como dito anteriormente na secção 3.2, os dados disponíveis para a

classificação são duas imagens *Landsat 5 TM*, uma datada de Julho de 2009 e outra de Novembro do mesmo ano. De cada imagem foram utilizadas todas as bandas excepto a banda térmica. A selecção dos atributos a definirem os objectos do espaço de classificação é um passo crítico no processo de classificação de uma imagem de satélite (Lu e Weng, 2004). Deste modo, é importante seleccionar apenas as classes que permitem a máxima separabilidade entre classes (Lu e Weng, 2004).

Para avaliar a separabilidade entre classes, recorreu-se à distância informativa de Jensen-Shannon. Assim, foram comparadas os quatro conjuntos de dados possíveis com os dados disponíveis, e para cada uma delas foi determinada da distância de Jensen-Shanon média entre subclasses espectrais. O critério utilizado para a selecção do conjunto de dados mais discriminativo consistiu na selecção do conjunto de dados que maximiza a distância de Jensen-Shannon média entre classes. O resultado pode ser observado na Tabela 7, onde o primeiro conjunto de dados é composto por apenas uma imagem Landsat (para o cálculo na tabela foi utilizada a imagem de Julho; no entanto, o resultado é análogo para a imagem de Novembro); o segundo é composto por uma imagem Landsat (Julho) e a sua respectiva banda sintética NDVI; o terceiro é composto por duas imagens Landsat (Julho e Novembro); e o quarto é composto por duas imagens Landsat e pelas respectivas bandas sintéticas NDVI.

Conjunto de dados	Dist. de Jensen-Shanon Média
Imagem Landsat (Julho)	22,77
Imagem Landsat (Julho) + NDVI (Julho)	89,83
Imagem (Julho) + Imagem (Novembro)	40,46

Imagem (Julho) + NDVI (Julho) + Imagem (Novembro) + NDVI (Novembro)	169,28

Tabela 7 – Distância de Jensen-Shanon média para cada conjunto de dados alternativo.

Os resultados mostram que o conjunto de dados que permite a maior separabilidade média entre classes é o conjunto de dados composto pelas duas imagens Landsat e as respectivas bandas sintéticas NDVI. Em particular, os resultados mostram ainda que a banda sintética NDVI introduz uma maior separabilidade entre classes que a introdução das seis bandas da segunda imagem Landsat. Deste modo, o conjunto de dados seleccionado para a classificação é composto por duas imagens Landsat (cada uma com seis bandas) mais as respectivas bandas sintéticas NDVI.

4.2 Nível de Confiança para o Conjunto de dados Seleccionado

Para a selecção do melhor nível de confiança, recorreu-se ao mesmo critério que anteriormente na selecção do melhor *conjunto de dados*. Os níveis de confiança possíveis foram: 90%, 95%, 97% e 99%. Apenas estes valores foram considerados uma vez que valores inferiores implicariam amostras se tornassem demasiadamente puras, o que tornaria o processo de generalização dos classificadores mais difícil (Webb, 2002; Kuncheva, 2004; Hastie et al., 2009). O gráfico na Figura 9 mostra o valor da distância de Jensen-Shanon média entre classes em função dos valores de significância do teste.

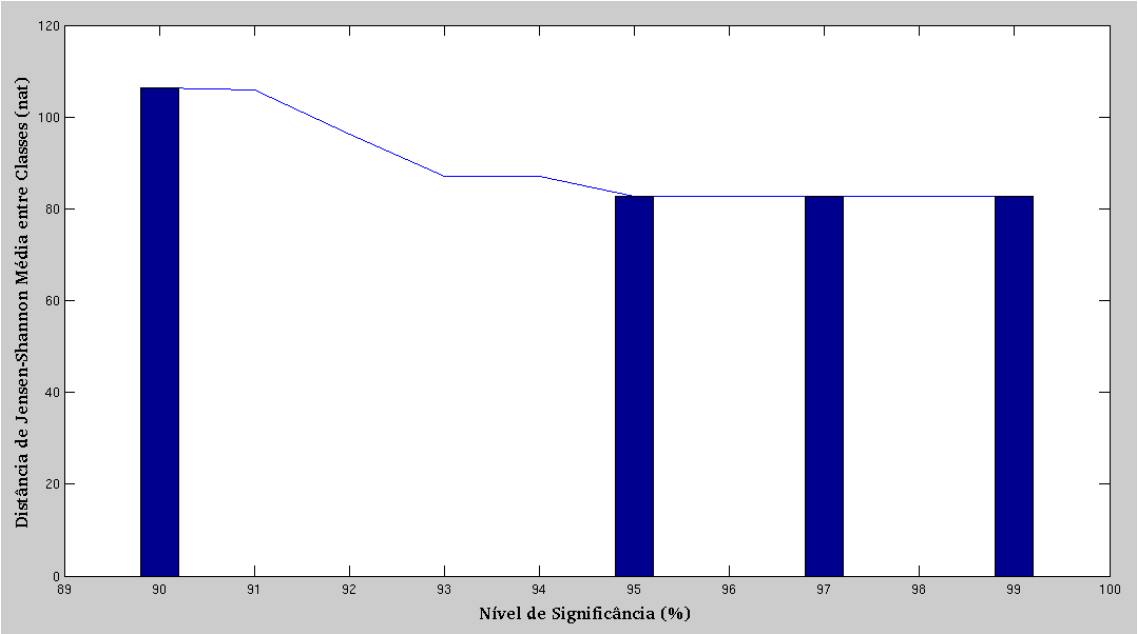

Figura 10 – Distância de Jensen-Shannon média entre classes em função do nível de significância do teste.

Os resultados mostram que o nível de significância que maximiza a distância de Jensen-Shannon média entre classes encontra-se nos 90%. Assim, a identificação estatística dos indivíduos do treino anómalos à classe é realizada com um nível de significância de 90%.

4.3 Avaliação da Qualidade dos Classificadores Simples

Como explicado na secção 3.5.4, os classificadores foram comparados por meio das curvas de aprendizagem, das curvas de robustez e da qualidade dos mapas que produzidos. Foram também tidos em conta os requisitos computacionais, ou seja, o intervalo de tempo necessário para a execução da classificação. Porem, antes de proceder à comparação dos classificadores, foi necessário fixar os parâmetros do classificador k-NN e das Árvores de Classificação. Para fixar estes parâmetros recorreu-se ao método *cross-validation* para identificar quais são os valores que maximizam a exactidão global. No caso do k-NN, o parâmetro a ser afinado é o número de vizinhos e no caso das Árvores de Classificação é o parâmetro que define o critério de paragem do

processo de divisão. Os resultados mostraram que o número de vizinhos que maximiza a exactidão global é 1 e que o valor para o critério de paragem é dado por 3. Assim, o classificador k-NN a ser utilizado nos testes será o 1-NN e nas Árvores de Classificação será aplicado um valor de *threshold* para a paragem da divisão de 3.

4.3.1 Curvas de Aprendizagem

As curvas de aprendizagem podem ser observadas na Figura 11.

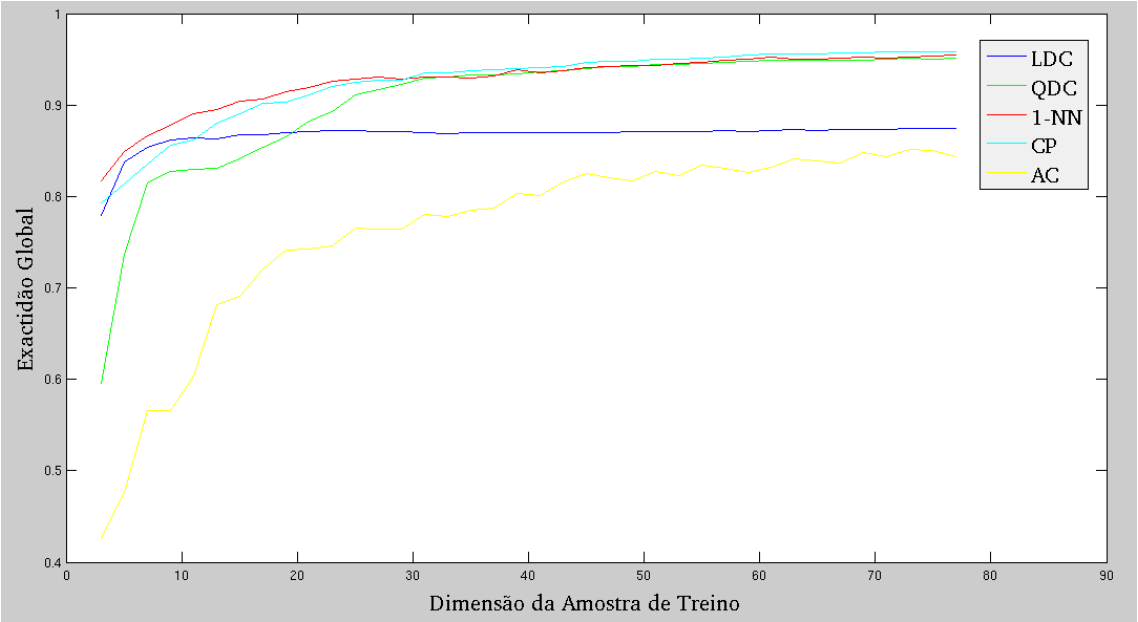

Figura 11 – Curvas de aprendizagem para os classificadores LDC, QDC, 1-NN, Classificador de Parzen (CP) e Árvores de Classificação (AC).

Os resultados mostram que cada classificador tem um número de mínimo de indivíduos de treino por classe necessário para obtenção da sua precisão global máxima possível, uma vez que a exactidão global tende a estabilizar após esse número. O LDC apresenta o menor número de indivíduos de treino por classe, 10, para uma precisão global máxima de cerca de 86% em *cross-validation;* os classificadores 1-NN e Parzen necessitam de cerca de 25 indivíduos de treino por classe para uma

precisão global máximo de 96%; o QDC precisa entre 30 a 40 elementos e, finalmente, as árvores de classificação necessitam mais de 60 indivíduos.

O facto do LDC necessitar de menos indivíduos de treino que os restantes classificadores pode ser explicado por dois factores.

Primeiro, o LDC é um classificador paramétrico, e portanto recorre à estimativa de um modelo para definir a informação do comportamento espectral de cada subclasse espectral. Este facto pode ser usado para explicar o melhor comportamento do LDC relativamente aos classificadores não paramétricos. Segundo, a facto do LDC necessitar de cerca de três vezes menos indivíduos de treino que o QDC pode ser explicada pela condição de homocedasticidade que é imposta para a construção do classificador.

De facto, no LDC, o que distingue uma classe de outra é o centro de massa da distribuição e não a variância, que é igual para todas as classes e é determinada pela matriz de covariâncias média. Assim, se existirem M classes e para cada classe existirem N indivíduos de treino, o algoritmo tem M x N indivíduos para estimar a matriz de covariâncias média. Por outro lado, o QDC necessita de estimar uma matriz de covariância especifica para cada classe, pelo que algoritmo possui apenas de N indivíduos para estimar as matrizes de covariâncias. Deste modo, o LDC consegue estimar uma matriz de covariância média suficientemente precisa com menos indivíduos por classe que o QDC, apesar de apresentar uma exactidão global inferior. Um outro modo de explicar a observação é dada por Hastie et al (2009). O algoritmo LDC necessita de estimar $K(K - 1)(d + 1)$ parâmetros para definir os hiperplanos de separação, onde K é o número de classes e d o número de dimensões do problema. O QDC, por outro lado, necessita de $K(K - 1)(d(d + 1)/2 + 1)$ parâmetros para a definição das suas fronteiras de discriminação. Assim, como o número de parâmetros do QDC é superior ao do LDC, também o número de indivíduos de treino terá que ser superior.

4.3.2 Curvas de Robustez

As curvas de robustez podem ser observadas na Figura 11.

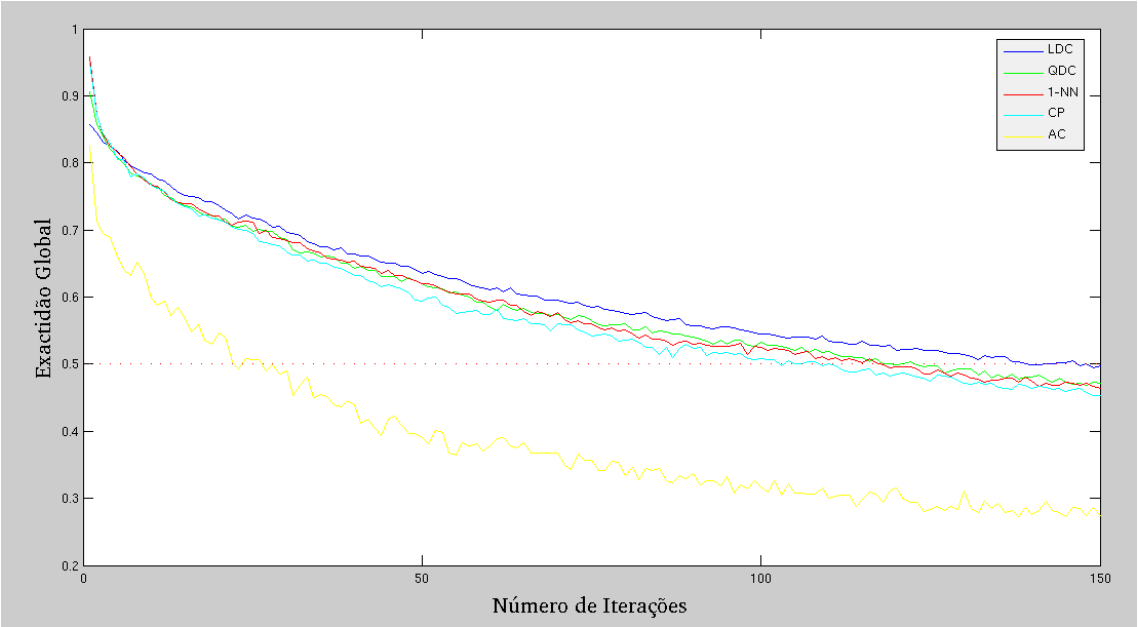

Figura 11 – Curvas de robustez para os classificadores LDC, QDC, 1-NN, Classificador de Parzen (CP) e Árvores de Classificação (AC).

Os resultados mostram que as Árvoeres de Classificação tendem a ser muito sensíveis à presença de ruído, uma vez que apresentam um decaimento mais rápido que os outros classificadores. Este resultado é coerente literatura, onde as árvores de classificação são qualificadas de classificadores instáveis (Kuncheva, 2004). O 1-NN, o classificador de Parzen e o QDC apresentam um comportamento semelhante, enquanto que o LDC é o classificador com um decaimento mais suave.

Os resultados obtidos podem ser explicados pela complexidade geométrica das fronteiras de discriminação. De um modo geral, um classificador com grande flexibilidade na definição de fronteiras de discriminação, tende a cair em *overtraining*, o que torna difícil a sua a adaptação objectos muito diferentes aos do treino (Kuncheva, 2004;

Hastie et al, 2009).

4.3.3 Comparação Experimental de Classificadores Simples

Em termos de tempo de execução, para uma imagem com 5990 x 3358 *pixels,* com uma profundidade de 8 *bits,* com 14 bandas e com uma amostra de treino de 6255 registos[8], os classificadores paramétricos e as Árvores de Classificação levaram entre 3 a 4 minutos para realizar a tarefa, enquanto que os classificadores não-paramétricos levaram mais de três horas. Estes resultados estão de acordo com a literatura, que reforça a ideia de que os classificadores não-paramétricos tendem a exigir grandes recursos computacionais e usualmente lentos (Webb, 2002; Kuncheva, 2004; Hastie et al, 2009). Portanto, do ponto de vista operacional, os classificadores paramétricos parecem ser a melhor solução.

Recorrendo ao protocolo de validação apresentado na secção 3.4.3.3, conclui-se que os classificadores testados apresentam resultados semelhantes como se pode observar na Tabela 8.

	LDC		QDC		1-NN		Parzen		AC	
Acrónimo	Util.	Prod.	Util.	Prod.	Util.	Prod.	Util.	Prod.	Util.	Prod.
1	86,7	100	95,2	100	87,5	100	90	100	100	100
2	89,6	65,2	89,3	68,5	91,8	67,2	89,5	58,6	92,1	56,5
3	97,9	94,2	98,9	90,5	93,2	97,2	94,5	97,2	91	96,1
4	95,2	99,3	93,3	96,9	95,3	92	97,7	94,5	93,9	95,6
5	82,2	94,6	78,6	93,6	83,1	82,2	82,6	83,5	81,3	93,8
6	100	95,5	100	76,5	96,3	89,7	90,6	90,6	88,9	84,2
8	50	25	80	88,9	77,8	63,6	85,7	85,7	70	63,6
10	100	100	100	100	77,6	100	79,2	100	100	100
E. G.	92		90,2		89		89,6		89,9	

Tabela 8 – Exactidão do utilizador, do produtor e global para cada classificador testado.

Como os resultados são muito semelhantes, aplicou-se o teste de McNemar (Tabela 9) a cada par de classificadores para verificar se as

[8]E numa máquina Linux com 4 GB de RAM e código Matlab.

diferenças eram significativas, em termos estatísticos, com um nível de confiança de 95%. Os testes mostram que não existem diferenças estatisticamente significativas entre os pares de classificadores avaliados, com excepção do par LDC / 1-NN.

LDC vs. QDC		
	QDC 1	QDC 0
LDC 1	435	25
LDC 0	16	24

LDC vs. 1-NN		
	1-NN 1	1-NN 0
LDC 1	430	30
LDC 0	15	25

LDC vs. CP		
	CP 1	CP 0
LDC 1	435	25
LDC 0	13	27

LDC vs. AC		
	AC 1	AC 0
LDC 1	433	27
LDC 0	16	24

QDC vs. 1-NN		
	1-NN 1	1-NN 0
QDC 1	422	29
QDC 0	23	26

QDC vs. CP		
	CP 1	CP 0
QDC 1	426	25
QDC 0	22	27

QDC vs. AC		
	AC 1	AC 0
QDC 1	432	19
QDC 0	17	32

1-NN vs. CP		
	CP 1	CP 0
1-NN 1	440	5
1-NN 0	8	47

1-NN vs. AC		
	AC 1	AC 0
1-NN 1	425	20
1-NN 0	24	31

CP vs. AC		
	AC 1	AC 0
1-NN 1	428	20
1-NN 0	21	31

Tabela 9 – Testes de McNemar para cada par de classificadores. LDC 1 = LDC certo; LDC 0 = LDC errado. Analogamente para os restantes classificadores.

Contudo, apesar das matrizes de erro mostrarem que os classificadores apresentam resultados superiores a 80% como desejado, contudo uma inspecção visual aos mapas mostra que existem diferenças significativas entre os classificadores e confusões que não são aceitáveis[9], como as

[9]No presente texto, uma confusão dir-se-á aceitável, se ocorrer entre subclasses espectrais da mesma

confusões entre classes de floresta e classes de agricultura. Por exemplo, o classificador 1-NN e o classificador de Parzen, apresentam mapas onde existe a confusão entre classes de agricultura e classes de floresta com a classe Água (Figura 12).

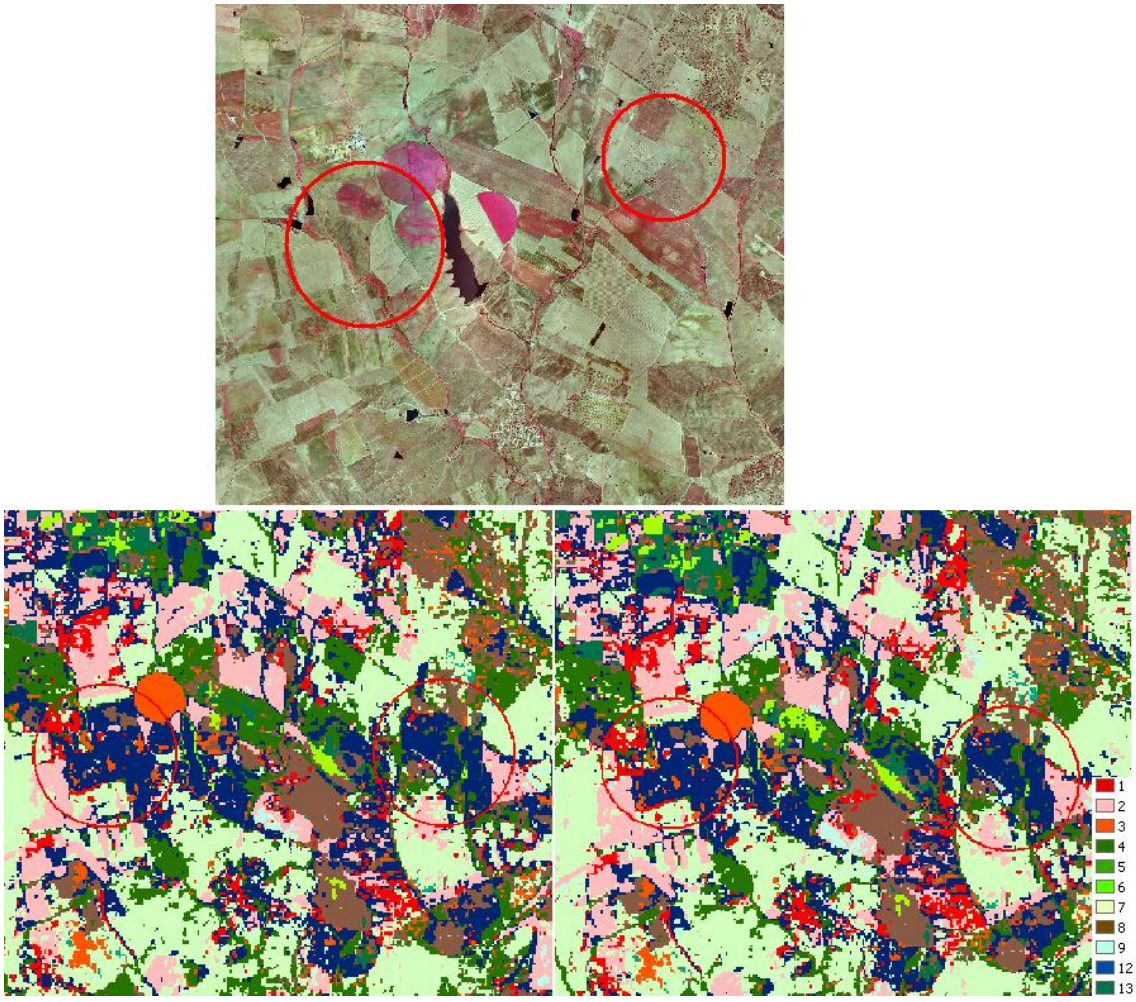

Figura 12 – Confusões entre a classe Água e as classes de Sequeiro e Regadio. Topo: Excerto de uma orto-imagem de uma zona da área de estudo. Esquerda: excerto do mapa produzido com o classificador 1-NN. Direita: excerto do mapa produzido com o classificador de Parzen. Os círculos a vermelho mostram dois exemplos da confusão entre a classe Água e as classes Sequeiro e Pastagem. A legenda da classificação está na nomenclatura das subclasses espectrais (Tabela 5).

classe da nomenclatura final.

No mapa produzido com o QDC salientam-se as confusões entre classes de floresta e classes de agricultura, em particular com a classe Regadio (Figura 13).

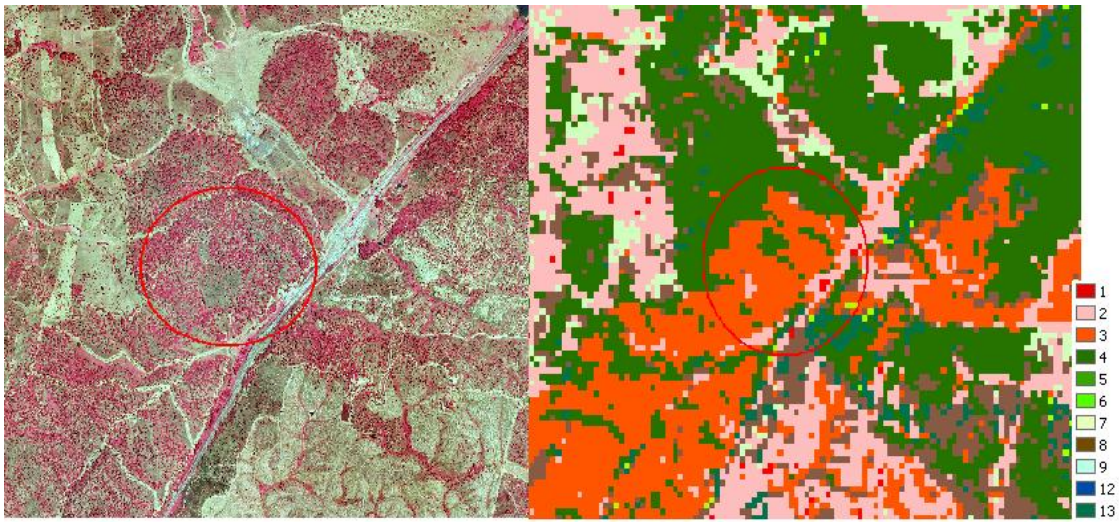

Figura 13 – Esquerda: excerto de uma orto-imagem da área de estudo. Direita: excerto do mapa produzido com o QDC. Os círculos a vermelho mostram um exemplo da confusão entre as classes de floresta e as classes de agricultura (neste exemplo, o regadio). A legenda classificação está na nomenclatura das subclasses espectrais (Tabela 5).

O mapa resultante das Árvores de Classificação apresenta o mesmo tipo de confusão que o mapa criado com QDC. Para mais, apresenta ainda uma forte comissão entre a classe Urbano e a classe Sequeiro e a classe Pastagem (Figura 14).

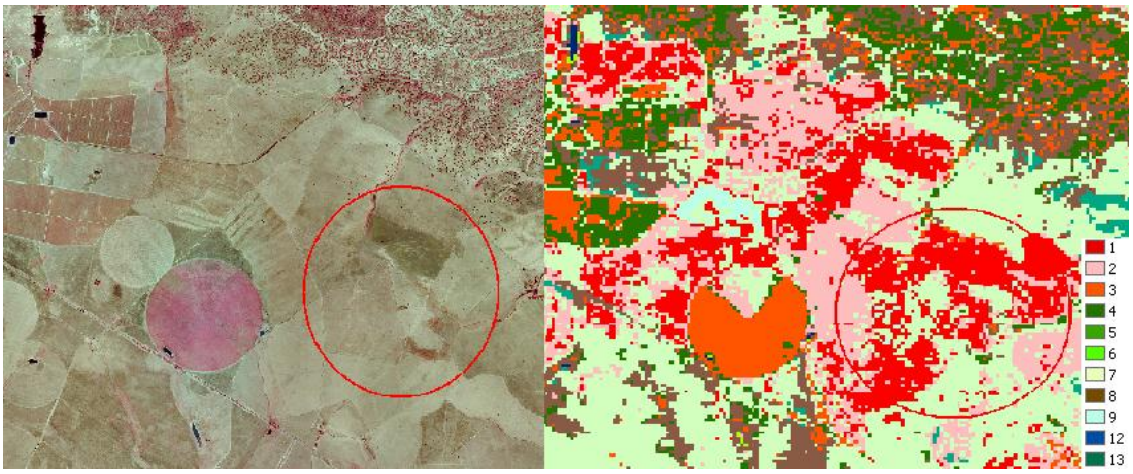

Figura 14 - Esquerda: excerto de uma orto-imagem da área de estudo. Direita: excerto do mapa criado com as Árvores de Classificação. O círculo a vermelho mostra um exemplo da comissão entre a classe Sequeiro e a classe Urbano. A legenda classificação está na nomenclatura das subclasses espectrais (Tabela 5).

A confusão entre as classes Sequeiro e Pastagem com a classe Urbano também é observada no mapa criado com o LDC (Figura 15). Contudo, este mapa apresenta uma estrutura paisagística mais coerente com os dados auxiliares, como o CLC06.

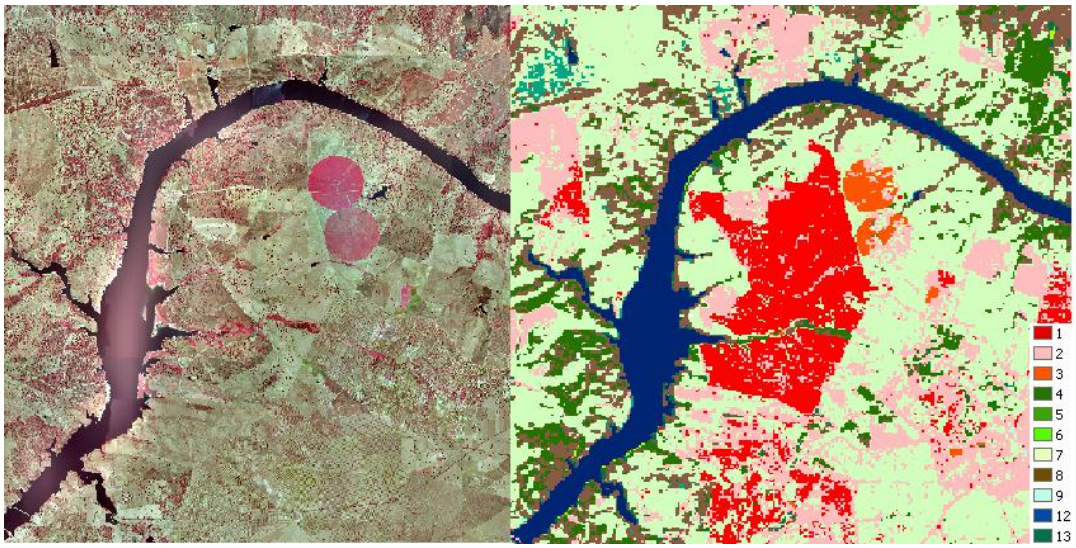

Figura 15 – Direita: excerto de uma orto-imagem da área de estudo. Esquerda: excerto do mapa produzido com o LDC. A legenda classificação está na nomenclatura das subclasses espectrais (Tabela 5).

Assim, dos classificadores analisados o classificador seleccionado foi o LDC. O classificador composto será baseado, portanto, no LDC. O objectivo deste classificador será mitigar as confusões observadas no mapa criado com o LDC singular, procurando manter as classificações correctamente realizadas pelo LDC.

4.4 Definição Especifica e Avaliação do Classificador Composto

Como dito na secção 3.6, o classificador composto recorre à divergência de Kullback-Leibler para identificar as classes mais prováveis de confusão com uma determinada classe inicial. A Tabela 10 foi construída recorrendo à equação 31 na secção 3.6.1:

$$D_{KL}(X,Y) = \frac{1}{2}\left\{\log\frac{|\Sigma_1|}{|\Sigma_0|} + tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_0^{-1}(\mu_1 - \mu_0) - N\right\}$$

onde | . | é a função determinante, tr (.) é a função traço, μ_0 é o vector médio da classe X, μ_1 é o vector médio da classe Y, Σ_0 é a matriz de

covariâncias da classe X, Σ_1 é a matriz de covariâncias da classe Y e N o número de dimensões do problema. As classes que se encontram nas linhas assumem a posição da variável X e as classes nas colunas a posição da variável Y.

	Urbano	Sequeiro	Regadio	Folhosas	Resinosas	Mistas	Pastagem	Matos	Solo Nu	Água	V. Esparsa
Urbano	0.00	0.40	5.48	4.50	21.84	10.55	0.52	4.67	0.92	100.00	4.78
Sequeiro	0.15	0.00	3.64	2.07	14.12	5.73	0.32	2.65	1.97	100.00	1.59
Regadio	2.78	2.67	0.00	0.76	3.02	1.13	1.47	0.96	9.46	100.00	7.86
Folhosas	11.71	10.57	1.39	0.00	0.09	0.27	2.33	0.31	17.90	100.00	0.68
Resinosas	28.06	25.60	5.50	0.41	0.00	0.33	7.67	1.03	40.48	100.00	1.67
Mistas	7.21	6.35	1.60	0.17	0.15	0.00	1.91	0.16	15.22	100.00	1.09
Pastagem	0.63	0.41	1.69	0.34	3.56	1.69	0.00	0.82	3.39	100.00	0.40
Matos	1.91	1.65	2.66	1.06	0.66	0.58	0.64	0.00	8.15	100.00	1.88
Solo Nu	0.34	1.21	6.76	19.19	51.95	28.26	2.07	7.30	0.00	100.00	16.26
Água	37.66	57.30	100.00	36.49	43.59	47.07	36.73	20.07	25.43	0.00	43.42
V. Esparsa	4.60	3.28	20.94	2.31	2.69	1.93	1.12	1.05	22.40	100.00	0.00

Tabela 10 - Matriz com as distâncias de Kullback-Leibler, normalizadas por linha por meio do valor máximo.

A partir da matriz das divergências de Kullback-Leibler podemos observar que todas as classes encontram-se extremamente afastadas da classe Água, uma vez que o máximo é assumido nessa classe. Este resultado é coerente com o que pode ser observado se for realizada uma redução de dimensões por meio da análise em componentes principais para duas dimensões (Figura 16).

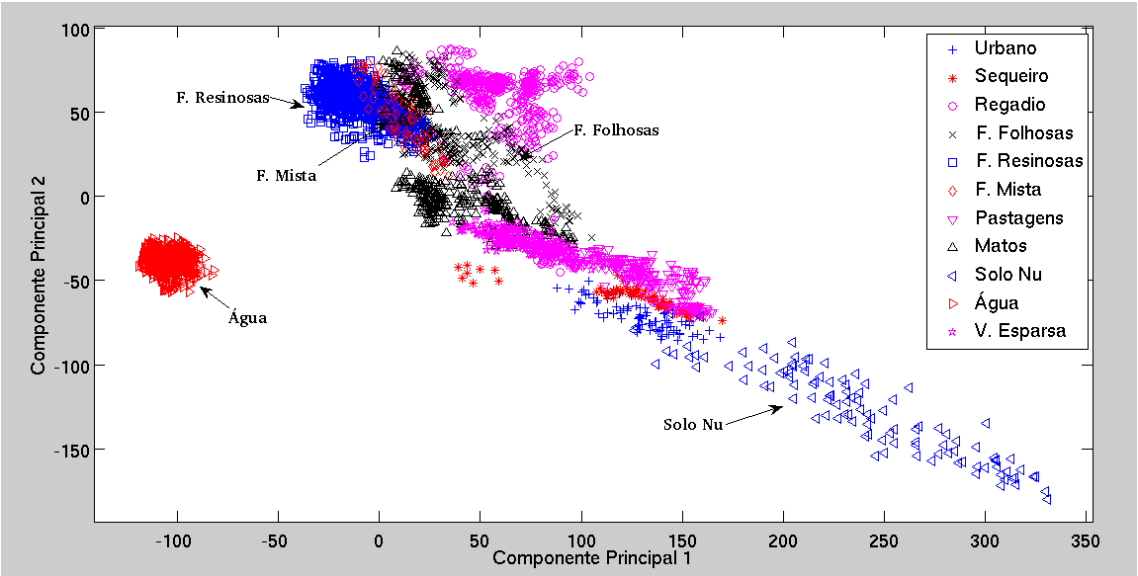

Figura 16 – Visualização da dispersão das classes no espaço de duas dimensões definido pelas duas primeiras componentes principais da transformação PCA.

Em particular, verificamos também que a classe Solo Nu se encontra muito afastada das classes de floresta, Floresta de Folhosas (19.19), Floresta de Resinosas (51.95) e Floresta Mista (28.26), o que faz sentido com o que é observado nas imagens, uma vez que a classe Solo Nu e as classes de floresta possuem respostas espectrais muito diferentes; de facto, estas classes encontram-se muito afastadas no espaço de classificação (Figura 16). Podemos ainda observar que, por exemplo, a classe Urbano encontra-se muito próxima das classes Sequeiro, Pastagem e Solo Nu, o que também é observado na Figura 17.

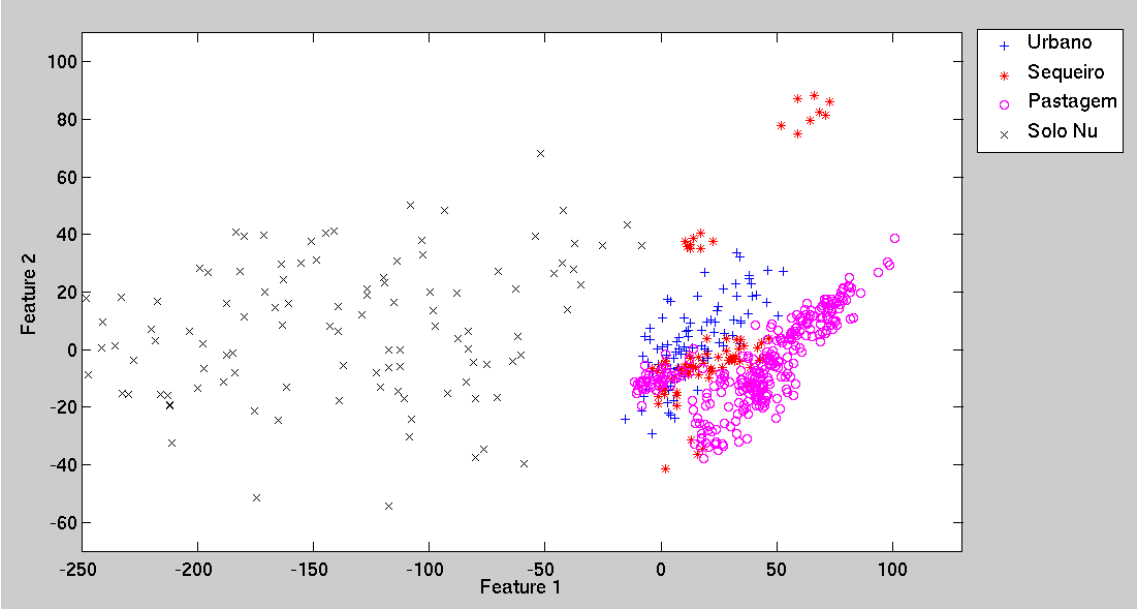

Figura 17 – Dispersão das classes Urbano, Sequeiro, Pastagem e Solo Nu no espaço de duas dimensões definido pela transformação PCA, que mostra a proximidade da classe Solo Nu das classes Urbano, Sequeiro e Pastagem, e sobreposição existente entre estas últimas classes.

Esta observação pode ser relacionada com a conclusão anteriormente

tirada de que os classificadores, em particular o LDC, tendiam a confundir as classes de Sequeiro, Pastagem e Solo Nu com a classe Urbano, ou seja, a classe Urbano tendia a dominar sobre estas classes. Mais especificamente, havia-se observado que o QDC confundia a classe Floresta de Folhosas com a classe Regadio. Esse facto pode ser observado na matriz das divergências de Kullback-Leibler: a classe Floresta de Folhosas é a classe que se encontra mais próxima da classe Regadio, o que sugere possível confusão entre estas duas classes.

Estas observações mostram que é possível utilizar a divergência de Kullback-Leibler como medida de proximidade entre classes espectrais de modo a identificar as classes que se encontram suficientemente próximas para serem confundidas umas com as outras por um classificador automático. Como a divergência de Kullback-Leibler é dependente da qualidade da amostra de treino, uma vez que são utilizadas as estimativas do vector médio e da matriz de covariâncias para a determinação desta medida, a proximidade detectada por esta medida pode advir da verdadeira semelhança espectral entre classes LULC ou de erros na selecção do treino. Portanto, a divergência de Kullback-Leibler pode ser utilizada para avaliar a qualidade de uma amostra de treino em termos de separabilidade entre classes.

O classificador composto, como dito na secção 3.6.3, será composto por dois classificadores simples dispostos sequencialmente. Como o LDC foi o classificador que apresentou os melhores resultados, o classificador será composto por dois LDC sequenciais. O primeiro LDC é igual ao classificador que foi testado, quer em termos algorítmicos quer em termos de treino. O segundo classificador possui o mesmo algoritmo que o primeiro, mas o treino é condicionado recorrendo a uma regra de selecção de classes, por meio da divergência de Kullback-Leibler.

Para a construção desta regra, observa-se a partir da matriz das divergências de Kullback-Leibler, que de um modo geral, para cada classe, existem entre quatro a cinco classes que se encontram mais

próximas, estando as restantes relativamente mais afastadas. Por exemplo, as classes mais próximas das classe Urbano são as classes Urbano (0), Sequeiro (0.40), Pastagem (0.52) e Solo Nu (0.92); as restantes classes possuem uma distância à classe Urbano superior a quatro unidades. Outro exemplo: as classes mais próximas da classe Floresta de Resinosas são Floresta de Resinosas (0), Floresta Mista (0.33), Floresta de Folhosas (0.41), Matos (1.03) e Vegetação Esparsa (1.67), estando as restantes classes afastadas por mais de cinco unidades. Assim, a regra utilizada para a selecção das classes a serem aplicadas no segundo classificador foi: se L foi o *label* atribuído ao *pixel* p pelo primeiro classificador, reclassificar p apenas com as quatro classes mais próximas da classe L. Verificou-se ainda que a classe Urbano tendia a dominar a classificação na fase de reclassificação de uma classe que não fosse Urbano. Esta facto pode ser explicado pela qualidade da amostra de treino recolhida para a classe Urbano. De facto, os urbanos da área de estudo são classificados na COS06 por "Urbanos Descontínuos", o que significa que são compostos por estruturas artificializadas intercaladas por ocupações agrícolas (sequeiro maioritariamente) ou de pastagem, que são precisamente as únicas classes em que a classe Urbano entra como alternativa na fase de reclassificação. Esta observação sugere que a amostra da classe Urbano encontra-se contaminada por indivíduos mal interpretados, apesar dos esforços para a identificação de *outliers*. Assim, a regra de selecção foi refinada de modo a não considerar a classe Urbano como uma alternativa viável na reclassificação de classes não Urbano, i.e. foi aplicada a seguinte regra de selecção: se L foi o label atribuído ao pixel p pelo primeiro classificador, reclassificar p considerando apenas as quatro classes mais próximas de L; se L for diferente de Urbano e se Urbano está nas quatro classes mais próximas, remover a classe Urbano e substitui-la pela quinta classe mais próxima.

Como este conjunto de classes é constituído pelas classes mais próximas, as fronteiras de discriminação tendem a possuir uma geometria muito complexa, devido à sobreposição entre classes. Assim,

para maximizar a separabilidade entre classes, foi aplicada uma transformação de Fisher. Os parâmetros da transformação de Fisher são estimados com os elementos da amostra de treino das classes determinadas pela regra. Como a aplicação de Fisher reduz a dimensionalidade do espaço, para um espaço de, no máximo, três dimensões[10], foram realizados testes para verificar qual o melhor número de dimensões para a classificação. Por inspecção visual aos mapas preliminares realizados, verificou-se que o melhor espaço para a classificação era definido com duas dimensões.

O resultado da validação do mapa produzido pelo classificador composto pode ser observado na Tabela 11.

	LDC		Class. Composto	
Acrónimo	Util.	Prod.	Util.	Prod.
1	86.7	100	100	75
2	89.6	65.2	92.5	76.6
3	97.9	94.2	97.4	97.4
4	95.2	99.3	98.5	99.3
5	82.2	94.6	87.3	97.4
6	100	95.5	100	100
8	50	25	66.7	40
10	100	100	100	100
E. G.	92		94.8	

Tabela 11 – Resultado da validação do mapa obtido com o classificador composto versus resultado da validação do mapa obtido com o LDC.

Os resultados mostram que, de um modo geral, há um ganho em termos de exactidão do produtor e do utilizador com excepção da classe Urbano que sofre uma redução na exactidão do produtor significativa. Esta redução pode ser explicada pela regra de selecção das classes que não favorece a reclassificação da classe Urbano em classes que não são Urbano. Na exactidão global existe um ganho de 2.8%. O

[10]Num problema com D dimensões e C classes, a transformação de Fisher define um espaço com um número de dimensões, no máximo, igual ao mínimo de C-1 e D. Como a regra restringe o número de classes para 4 e o número de dimensões é 14, então o novo espaço de classificação terá, no máximo, 3 dimensões.

teste de McNemar mostra de que as diferenças são significativas (Tabela 12), pelo que os resultados obtidos pelo classificador composto são globalmente melhores.

Classificador Composto vs. LDC		
	LDC 1	**LDC 0**
CC 1	459	15
CC 0	1	25

Tabela 12 – Teste de McNemar, Classificador Composto (CC) vs. LDC.

Visualmente, podemos observar que existe um ganho na coerência paisagística (Figura 18).

Figura 18 – Correcção dos erros de classificação na classe Urbano. Centro: Orto-imagem de uma parte da área de estudo. Esquerda: excerto do mapa produzido com o LDC onde se pode ver uma grande mancha de Urbano (vermelho). Direita: excerto do mapa produzido com o classificador composto onde se pode observar que os *pixels* anteriormente classificados como Urbano são agora classificados como Sequeiro (Rosa) e Pastagem (verde muito claro).

Na figura 19, podemos observar o efeito da reclassificação nas áreas urbanas. Observa-se que as manchas anteriormente classificadas como Urbano perdem alguma extensão. Este efeito é apenas devido à reclassificação e não à restrição imposta pela regra de selecção sobre a classe Urbano, uma vez que, sempre que o primeiro classificador

identifica um *pixel* como Urbano, essa classe é incluída na reclassificação

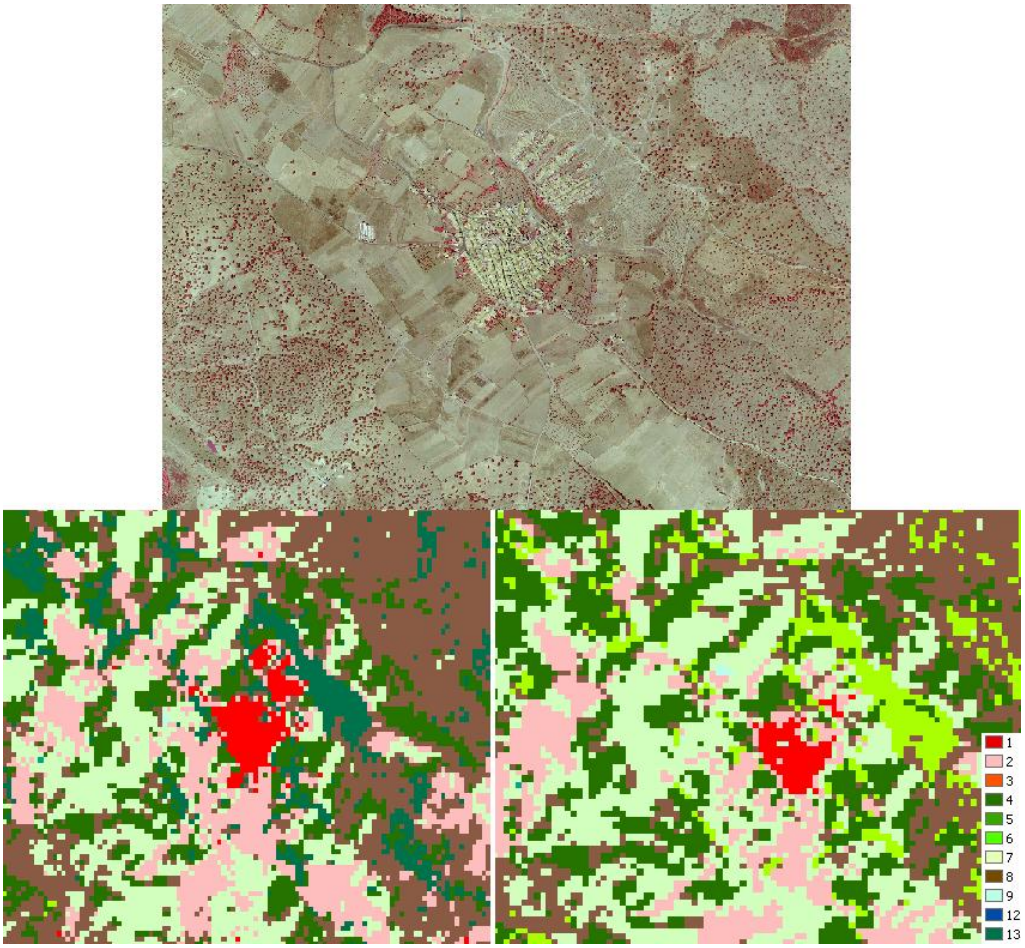

Figura 19 – Acção da regra de selecção sobre a classe Urbano. Topo: orto-imagem de uma parte da área de estudo. Esquerda: excerto do mapa produzido pelo LDC. Direita: excerto do mapa produzido pelo classificador composto. A legenda da classificação está na nomenclatura das subclasses espectrais (Figura 5).

5 Conclusão

O presente estudo foi realizado no contexto do projecto DWE que tem como objectivo principal o desenvolvimento e produção de mapas de indicadores de desertificação a partir de imagens de satélite, através de

uma aplicação informática a ser desenvolvida no decurso do projecto. Para isso foram comparados vários classificadores (LDC, QDC, 1-NN, Classificador de Parzen e as Árvores de Classificação) para identificar aquele que melhor se ajustava às condições impostas no projecto.

Os resultados sugerem que o LDC é o classificador que necessita de menos treino para alcançar a sua exactidão global máxima e mais resistente ao ruído. Em termos da qualidade dos mapas produzidos, os classificadores comportaram-se, de um modo geral, de forma similar, sendo a precisão global mínima de 89%. No entanto, o classificador que mostrou produzir um mapa mais coerente com a informação auxiliar sobre a área de estudo foi o LDC. Deste modo, o LDC foi o classificador seleccionado. Contudo, e apesar da validação do mapa apresentar exactidões globais excelentes, uma inspecção visual ao mapa mostrou que existiam confusões inaceitáveis, como a confusão da classe Sequeiro com a classe Urbano. Deste modo, foi desenvolvido um classificador composto para mitigar as confusões encontradas.

O classificador composto desenvolvido neste trabalho recorre a medidas de proximidade informativa, baseadas na medida de entropia informativa de Shannon, para identificar as classes mais próximas de uma determinada classe. Estas medidas mostraram conseguir identificar as classes que se encontram mais próximas umas das outras no espaço espectral, conseguindo mesmo explicar alguns dos erros de classificação que se verificavam nos mapas produzidos.

O classificador composto foi construído utilizando o algoritmo do LDC como base. O conceito subjacente à sua aplicação consiste num ciclo de classificação e reclassificação aplicado a cada elemento do espaço de classificação; no presente estudo, os *pixels* de uma imagens de satélite. Na fase de classificação, o *pixel* a ser classificado é avaliado pelo LDC, com a amostra de treino inicialmente recolhida para a classificação das imagens de satélite, ao qual é atribuído um *label* de uma das classes da nomenclatura inicial. Seguidamente, o *pixel* é reclassificado com o mesmo algoritmo, mas recorrendo a uma amostra de treino composta

apenas pelas classes mais próximas da classe atribuída na fase de classificação anterior. Essas classes são seleccionadas recorrendo a uma regra que depende dos valores da divergência de Kullback-Leibler entre classes. A regra que, para a presente área de estudo, produziu melhores resultados consistiu na selecção das quatro classes mais próximas da classe atribuída na fase de classificação, com um tratamento especial no caso da classe Urbano se encontrar incluída na lista das classes mais próximas. Seguidamente, como as classes seleccionadas se encontram muito próximas umas das outras, dificultando a classificação, foi necessário aplicar uma transformação de Fisher para maximizar a separabilidade entre classes. O último passo no algoritmo do classificador composto é a reclassificação do *pixel*, que é realizada no novo espaço de classificação definido pela transformação de Fisher e utilizando apenas as classes seleccionadas com classes possíveis de atribuição recorrendo ao algoritmo do LDC. Este segundo classificador mostrou melhorar os resultados obtidos com o classificador simples LDC, resolvendo uma grande parte dos erros de classificação encontrados no mapa produzido pelo LDC.

Concluindo, o presente estudo conseguiu mostrar que a combinação de um classificador simples, como o LDC, composto num esquema sequencial de classificação e recorrendo à medida de proximidade informativa de Kullback-Leibler, permite melhorar os resultados obtidos com um classificador singular.

6 Referências

Aronoff A., 1982. Classification Accuracy: User Approach. Photogrammetric Engeneering and Remote Sensing, Vol. 48, Nº 8, Agosto 1982, pp. 1299-1307.

Aronoff A., 1985. Classification Accuracy: User Approach. Photogrammetric Engeneering and Remote Sensing, Vol. 51, no. 1, Janeiro 1985, pp. 99-111.

Akgun A., Eronat A., Turk N., 2009. Comparing Different Satellite Image Classification Methods: An Application in Ayvalik District, Western Turkey. XXXV ISPR Congress Proceedings. http://www.isprs.org/proceedings/XXXV/congress/comm4/papers/505.pdf acedido a 28 de Outubro de 2010.

Conese C., e Maselli F., 1994. Evaluation of contextual, per-pixel and mixed classification procedures applied to a subtropical landscape. Remote Sensing Reviews, no.9, pp. 175-186.

Chen C. M., e Stow, D.A., 2002. The effect of training strategies on supervised classification at different spatial resolution. Photogrammetric Engineering and Remote Sensing, no. 68, pp. 1115-1162.

Cingolani, A.M., Renison, D., Zak, M.R. and Cabido, M.R., 2004, Mapping vegetation in a heterogeneous mountain rangeland using Landsat data: an alternative method to define and classify land-cover units. Remote Sensing of Environment, no. 92, pp. 84–97.

Congalton, R.G. e Green, K., 1999, Assessing the Accuracy of Remotely Sensed Data:Principles and practices. New York: Lewis Publishers.

Chilar, J., Xiao, Q., Chen, J., Beaubien, J., Fung, K. and Latifovic, R., 1998. Classification by progressive generalization: a new automated methodology for remote sensing multispectral data. International Journal of Remote Sensing, no.19, pp.2685–2704.

Cover T., e Thomas J., 2006. Elements of Information Theory. Second Edition. Wiley Inter-Science, John Wiley and Sons, Inc.

Chang C., 2000. An Information-Theoretic Approach to Spectral Variability, Similarity, and Discrimination for Hyperspectral Image Analysis. IEEE Transactions on information theory, vol. 46, no. 5, August 2000.

Carrão, H., A. Araújo, P. Gonçalves, and M. Caetano, 2010. Multitemporal MERIS images for land cover mapping at national scale: the case study of Portugal. International Journal of Remote Sensing, Vol. 31, no. 8, pp 2063-2082.

Conchran W. G., 1977. Sampling Techniques. Second Edition. John Wiley and Sons, Inc.

Dicks, S. and Lo, T., 1990. Evaluation of thematic map accuracy in a land-use and land-cover mapping program. Photogrammetric Engineering and Remote Sensing, Vol. 56, no. 9, pp. 1247-1252.

Fukunaga K., 1990. Introduction to Statistical Pattern Recognition. Second Edition. Morgan Kaufmann Academic Press.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment. No. 80, pp. 185–201.

Ginevan M. E., 1979. Testing Land-Use Map Accuracy: Another Look. Photogrammetric Engeneering and Remote Sensing, Vol. 45, Nº 10, Outubro 1979, pp. 1371-1377.

Hastie T., Tibshirani R. e Friedman J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer Series in Statistics.

Hughes G., 1964. On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Transactions on Information Theory, vol. IT-14, no 1, January 1964.

Johnson, R.A. e Wichern, D.W., 1998. *Applied Multivariate Statistical Analysis*. 4th Edition. New Jersey: Prentice Hall, Upper Saddle River.

Kuncheva L., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience. John Wiley and Sons, Inc.

Kullback S. e Leibler R. A., 1951. On Information and Sufficiency. The Annals of Mathematical Statisitics, www.Jstor.org. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms/1177729694 Acedido a 28 de Outubro de 2010.

LU, D. andWENG, Q., 2004. Spectral mixture analysis of the urban landscapes in Indianapolis with Landsat ETM+ imagery. Photogrammetric Engineering and Remote Sensing, no. 70, pp. 1053–1062.

LO, C.P. and CHOI, J., 2004, A hybrid approach to urban land use/cover mapping using Landsat 7 Enhanced Thematic Mapper Plus (ETM +) images. International Journal of Remote Sensing, no. 25, pp. 2687–2700.

Mather P., 2004. Computer Processing Remotely Sensed Images: An Introduction. Third Edition. John Wiley and Sons, Inc.

Pestana D. e Velosa S., 2002. Introdução à Probabilidade e à Estatistica Volume 1. Fundação Calouste Gulbenkian, Lisboa.

Park S. C. Y, Lawrence K. C. e Windham W. R., 2007. Fisher Discriminant Analysis for Improving Fecal Detection Accuracy with Hyperspectral

Images. American Society of Agriculture and Biological Engineers. Transactions of the ASABE, Vol. 50, no 6, 2275-2283.

Panigada C., Fava F., Zucca C., 2009. Remote Sensing and Land Degradatin in Tropic Drylands, a Review. Nucleo Ricerca Desertificazione Università di Sassari.

Ripado, M. F., 1991. Calendário Rural, Lisboa: Litexa Editora.

Stehman, S.V. and Czaplewski, R.L., 2003. Introduction to special issue on map accuracy. Environmental and Ecological Statistics, no. 10, pp. 301–308.

Stehman, S.V. and Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. Remote Sensing of Environment, no. 64, pp. 331–344.

Shannon C. E., 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal,* Vol. 27, pp. 379–423, 623–656, July, October, 1948.

Wilkinson G. G., 2005. Results and Implications of a Study of Fifteen Years of Satellite Image Classification Experiments. IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, March 2005.

Warrender, C.E. and Augusthein, M.F., 1999, Fusion of image classification using Bayesian techniques with Markov random fields. International Journal of Remote Sensing, no. 20, pp. 1987–2002.

Webb A., 2002. Statistical Pattern Recognition. Second Edition. John Wiley and Sons, Inc.

Wulder, M., Franklin, S., White, J., Linkes, J., and Magnussen, S., 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. International Journal of Remote Sensing no. 27, pp. 663-683.

Wickham, J.D., Stehman, S.V., Smith, J.H. and Yang, L., 2004. Thematic accuracy of the 1992 National Land-Cover Data for the western United States. Remote Sensing of Environment, no. 91, pp. 452–468.

Zhu, Z., Yang, L., Stehman, S.V. and Czaplewski, R.L., 2000. Accuracy assessment from the US Geological Survey Regional Land Cover Mapping Program: New York and New Jersey Region. Photogrammetric Engineering and Remote Sensing, no.66, pp. 1425–1435.

Zhou S. K. e Chellapa R., 2004. Kullback-Leibler Distance between Two Gaussian Densities in Reproducing Kernel Hilbert Space. Conference Proceeding in ISIT 2004, Chicago, DC, June 27-July 2, 2004.